The Major Transitions in the Evolution of Language

Willem H. Zuidema, MSc.

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to

Theoretical and Applied Linguistics
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

August 2005

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

"It is astonishing what language can do. With a few syllables it can express an incalculable number of thought, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thoughts corresponding to the parts of a sentence, so that the structure of the sentence serves as the image of the structure of the thoughts." (Frege, 1923)

"A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue." (Darwin, 1871, p. 91)

"Humans and chimpanzees are nevertheless very similar in their proteins, on the average, but vastly different in the sizes of their brains and their ability to write books about each other." (Lewontin, 1998, p. 117)

Abstract

The origins of human language, with its extraordinarily complex structure and multitude of functions, remains among the most challenging problems for evolutionary biology and the cognitive sciences. Although many will agree progress on this issue would have important consequences for linguistic theory, many remain sceptical about whether the topic is amenable to rigorous, scientific research at all. Complementing recent developments toward better empirical validation, this thesis explores how formal models from both linguistics and evolutionary biology can help to constrain the many theories and scenarios in this field.

I first review a number of foundational mathematical models from three branches of evolutionary biology – population genetics, evolutionary game theory and social evolution theory – and discuss the relation between them. This discussion yields a list of ten requirements on evolutionary scenarios for language, and highlights the assumptions implicit in the various formalisms. I then look in more details at one specific step-by-step scenario, proposed by Ray Jackendoff, and consider the linguistic formalisms that could be used to characterise the evolutionary transitions from one stage to the next. I conclude from this review that the main challenges in evolutionary linguistics are to explain how three major linguistic innovations – combinatorial phonology, compositional semantics and hierarchical phrase-structure – could have spread through a population where they are initially rare.

In the second part of the thesis, I critically evaluate some existing formal models of each of these *major transitions* and present three novel alternatives. In an abstract model of the evolution of speech sounds (viewed as trajectories through an acoustic space), I show that combinatorial phonology is a solution for robustness against noise and the only evolutionary stable strategy (ESS). In a model of the evolution of simple lexicons in a noisy environment, I show that the optimal lexicon uses a structured mapping from meanings to sounds, providing a rudimentary compositional semantics. Lexicons with this property are also ESS's. Finally, in a model of the evolution and acquisition of context-free grammars, I evaluate the conditions under which

hierarchical phrase-structure will be favoured by natural selection, or will be the outcome of a process of cultural evolution.

In the last chapter of the thesis, I discuss the implications of these models for the debates in linguistics on innateness and learnability, and on the nature of language universals. A mainly negative point to make is that formal learnability results cannot be used as evidence for an innate, language-specific specialisation for language. A positive point is that with the evolutionary models of language, we can begin to understand how universal properties and tendencies in natural languages can result from the intricate interaction between innate learning biases and a process of cultural evolution over many generations.

Acknowledgements

The seven chapters of this thesis are based on research I did in six different institutes[1], in five different cities, in four different countries, using at least three different methodologies[2], drawing mainly on work from two different fields[3] but all focused on one single topic: the evolutionary origins of human language. I feel I have been very fortunate during this research, having found wonderful colleagues to work with in each of these institutes, and having made great friends in each of the cities I lived in.

Although it may not seem like it, I do in fact very much dislike moving. Leaving friends, colleagues, apartments, habits and favourite places behind has at each transition been painful. The move from Brussels to Edinburgh has no doubt been the worst. I was really fond of my little balcony on the *Sint Katelijneplein*, of the *wittekes* in Café Kafka and *frites avec sauce Andalouse* in snackbar Aquarium, of my desk on the tenth floor of *gebouw F* and the Friday night Leffes in the Kultuur Kafé, of the great food in restaurants across Brussels, of my feeble attempts to speak and understand Flemish and French, and most of all, of my friends and colleagues there. When it became clear that I had to leave Brussels, there was only one place that could be an acceptable alternative: Edinburgh, a city already on the top of my list of favourite cities in the world, with a university where I could continue the research I wanted to do. Nevertheless, I must admit I left Brussels with a heavy heart.

Looking back, the move to Edinburgh has been the proverbial blessing in disguise. I quickly discovered the many things Edinburgh and Scotland have to offer – from cosy pubs to snowy mountains. But most importantly, I found the best academic climate I could wish for. All those seminars, journal clubs, lab meetings and top class lectures – far too many to attend them all, but exactly what I had imagined doctoral research to be like. I sat in on many postgraduate lectures in

[1] Theoretical Biology, Utrecht University; Sony Computer Science Laboratory – Paris; Artificial Intelligence lab, Free University Brussels; Language Evolution and Computation research unit, University of Edinburgh; Institute for Cell, Animal and Population Biology, University of Edinburgh; Institute for Logic, Language and Computation, University of Amsterdam.

[2] Verbal reasoning, mathematical modelling and computer simulation.

[3] Linguistics and evolutionary biology.

population biology and theoretical linguistics, and learnt many things I wish I had known about many years earlier. I have enjoyed immensely the countless discussions about my work and that of others – sometimes in formal supervisions, but mostly in an informal setting near the Population Genetics coffee table or in the Linguistics Common Room.

I am extremely grateful to the people that made this possible. First and foremost, my supervisors **Jim Hurford**, **Simon Kirby** and **Nick Barton**. I first contacted Jim when things started to look grim in Brussels. He has been extremely supportive and welcoming from the start, even when it was still far from certain I would actually move to Edinburgh. Whilst I was in Edinburgh, I found Jim to be even more original and knowledgeable than I had expected. Free from politics and convention that spoil so much of both science and *joie de vivre*, Jim is an example for how to be successful, happy and liked in academia.

Simon, my principal supervisor, quickly convinced me that Edinburgh was the best place for me to do my PhD research, as he had in fact already told me in 1999. He enthusiastically let me continue the research I had started in Brussels, but at crucial moments encouraged me to learn more about topics that I had neglected before. His own work continues to be an inspiration for my research, and up until the final phases of this thesis I continued to discover relevant new insights in it.

Nick must be the world's most open-minded mathematical biologist. I emailed him in the summer of 2002 – unaware of his work and his fame in evolutionary biology – to enquire about a fellowship in *quantitative genetics* for which the application deadline had already passed. This was a desperate move of an artificial intelligence student heading for financial disaster. I hadn't expected that he would respond as positively and helpfully as he did, let alone that this would eventually have such a profound influence on my research. Not only did he, to my amazement, tolerate a "linguist" – ignorant about both population genetics and *Mathematica* – in his lab, but he was actually frequently available for pointed advise and a refreshing informed outsider's perspective. He patiently encouraged me to read up on the classic literature in evolutionary biology (as reflected in chapter 2) and to make my ideas about language evolution more precise, formal and testable. The result was a lot of extra work, but it has been extremely rewarding for me and I hope it has made the thesis better.

I am grateful also to **Paulien Hogeweg** and **Gert Westermann** for writing letters of recommendation under intense time pressure. Paulien was the supervisor of my Master's research in Utrecht, and taught me a lot about biology, modelling and the ethics of academic research. She

continued to give valuable advise while I was in Paris and Brussels. Joint work and fruitful discussions with her formed the basis of the research described in chapter 6. Gert became a good friend while I was in Paris, where we both longed for returning to a university environment. His knowledge and excitement about cognitive science had a major influence on me; joint work that we started already in Paris eventually led to the research described in chapter 5.

Many other people have influenced my research – too many to list them all. But four of them have been particularly important. I have immensely enjoyed the discussions and brainstorm sessions with **Joachim de Beule** during my two years in Brussels. He has had a major influence in rekindling my interest in "classical" artificial intelligence. I still hope we'll have a chance one day to pick up on our fruitful collaborations. **Bart de Boer**'s broad knowledge about linguistics and phonology has been inspiring. Joint work with him that started during a train ride from Switzerland back to Brussels, has eventually led to the research described in chapter 4.

Andy Gardner's enthusiasm about mathematical modelling in general, and about the Price Equation in particular has been contagious. He was one of the main reasons why I enjoyed being at ICAPB so much, and I hope there will be opportunities to work together again in the future. **Tim O'Donnell** has had such an influence on my research, that I had almost included him in my list of supervisors, even though officially he hasn't even started his own PhD research yet. Joint work with him is reflected in chapter 6. Tim's knowledge of the mathematical linguistics and game theory literature has been imposing at times. His pages long critiques of my papers and chapters have sometimes made me despair, but they always raised so many new and interesting issues that I always continued to look forward to work with him again and learn more. Eventually, something brilliant will come out of it.

I have received so many detailed and useful comments on the chapters of this thesis from various people, that I haven't always been able to follow up on the excellent suggestions. In addition to those of my supervisors Simon, Jim and Nick, the most extensive comments were from Andy on chapter 2, Tim on chapters 3, 5, 6 and 7, Bart and Matina Donaldson on chapter 4, Joachim on chapter 6 and Marian Counihan on chapter 7. Kenny Smith and Anna Parker have proofread many versions of the papers on which these chapters are based. Many thanks for the help; I'd be happy to return the favour, and will try to be equally critical.

PhD students do need to pay the bills, as is sometimes forgotten. I am grateful for the funding I received from Sony CSL-Paris, from the GOA grant "Origins of language and meaning" of the Vrije Universiteit Brussel and the Flemish Government, from a "Cultuurfondsbeurs" of the Prins

Bernhard Cultuurfonds in Amsterdam and from a Marie Curie fellowship of the European Commision at the Institute for Cell, Animal and Population Biology of the University of Edinburgh. I also thank the VUB, NIPS, The University of Edinburgh Development Trust and ICAPB for providing funding for attending various conferences.

Finally, I have been very fortunate to have met so many wonderful colleagues and friends abroad, whilst continuing to be able to count on my friends and family back home. Thanks to Aukje, Diederik, Minke, Peter, Mark, Sandra, Martijn, Auke, Ivo, Rudmer, Alex, Maartje, Marinus, Onno, Lotte, Anneroos, Stan, Edwin, my new colleagues and others in the Netherlands; thanks to Regina, Arnd, Thessa, Joachim, Veronique, Tony, Sarah, Bart, Cecile, Bart, Barbara, Joris, Bart, Dominique, Carlos Ruben and others in Brussels; thanks to Greg, Benjamin, Hilke, Inga, Annemieke, Francesca, Amy, Malvina, Viktor, Dan, Monica, Andrew, Henry, Hajime, Carrie, Christine, Marisa, Dan, Linda, Penny, Angeles, Angus, Mathieu and others in Edinburgh, and to Eric in lots of different places. A special thanks to my parents, Theo and Annemarie, who encouraged me to wonder about life and the universe from a very early age. I wish my grandparents had lived long enough to see the result of my years abroad, during which I could visit them much less frequently than I would have liked. I miss them much.

Contents

CHAPTER 1

General Introduction

1.1 Why Study the Evolution of Language?

Language evolution is a booming field, there can be no doubt about it. Christiansen & Kirby (2003c) counted 94 published papers per year in the period 1990–2002 in the on-line database "ISI Web of Science". *Science*, *Nature* and other high-profile journals publish many papers per year on the topic. There is a biennial conference, which had its fifth edition in March 2004, numerous workshops, and a book-series by Oxford University Press. Each year, collections of academic papers on language evolution are published (Hurford, Studdert-Kennedy & Knight, 1998; Knight, Hurford & Studdert-Kennedy, 2000; Briscoe, 2002b; Wray, 2002; Cangelosi & Parisi, 2002; Christiansen & Kirby, 2003a; Tallerman, 2004), as well as numerous popular science articles and books. There are funding opportunities earmarked for language evolution research, specialised research groups, and a large number of university courses.

This surge in interest followed – coincidence or not – the publication of the most cited[1] paper from the field, Steven Pinker and Paul Bloom's position paper *Natural Language and Natural Selection* (1990), and Pinker's bestselling popular science book *The language instinct* (1994) expanding the argument from that paper. Pinker and Bloom argued that there is every reason to believe the human "language instinct" originates in a process of classical, Darwinian evolution, and their argument apparently hit a chord.

It is not difficult to understand why so many researchers are interested in the origins of language. After all, language is a defining characteristic of our own species, and a sine qua non for human society, religion, culture, technology and, indeed, science. Its origins are fascinating in

[1] ISI Web of Science lists 257 citing articles, more than any other citation hit in language evolution I know about; the on-line *Language Evolution and Computation Bibliography and Resources* (Wang, 2004) has 79 citing papers in its database, more than any of the other 757 papers in the database.

their own right and, moreover, a better understanding of language origins is likely to have major implications for theories of the nature, use and acquisition of language and, perhaps, for the study of animal communication. Pinker & Bloom (1990) argued that human language's unique features can be understood as having evolved for the purpose of conveying an unbounded number of messages over a limited, linear channel.

However, Pinker and Bloom's paper – although important in countering an anti-evolutionary stance of many linguists at the time – is a peculiar paper to be at the heart of the field. The paper does not present a real theory, other than the proposal that we should think about language as an "adaptation". Crucial components of evolutionary explanations – the variation available for evolution, the intermediate steps, the selection pressures moving our ancestors from one stage to the next – are missing. Rather, as evolutionary biologist Richard Lewontin (1990) points out, the paper adopts an argument that is most popular with *critics* of the theory of Natural Selection: the Argument from Design. Because language is too complex to have arisen as a side-effect, Pinker & Bloom argue, there are no coherent alternatives to a classical Darwinian explanation. Except for a sketchy survey of factors that could have played a role – information sharing, sexual selection, the Baldwin effect – the paper does not even start with providing a candidate evolutionary scenario.

Many others in the field have presented more substantive theories of language evolution. However, in the hundreds of papers that have appeared since 1990, no real consensus has emerged about the fundamental questions of the field. Christiansen & Kirby (2003b), in a review of consensus and controversies, list only methodological issues as points of emerging consensus: the need for interdisciplinary research, the need for formal modelling and the need to investigate possible precursors of the language faculty in non-human animals and prelinguistic hominids. Controversies, in contrast, abound. Researchers in language evolution are sharply critical of each other's work. Almost all chapters in the recent collection edited by Christiansen & Kirby (2003a) start with criticising fundamental misunderstandings and omissions in the field as a whole. For instance:

- Newmeyer (2003) complains about the limited involvement of linguists and linguistic theory in theorising about language origins;
- Bickerton (2003b) agrees, but adds that ignorance of linguistics is a special case "of a much more widespread tendency [...]. All too often, a writer whose home is in one or other of [the relevant] disciplines will make a proposal that is unacceptable in terms of one or more of the other relevant disciplines". Bickerton is particularly concerned about the lack of interest in

the evolution of complex syntax, and in relating linguistic innovations with major cultural changes in hominid evolution;

- Lieberman (2003), on the other hand, complains about the lack of interest in the articulatory and acoustic constraints on reaching the very high rate of information transfer in human speech. With Hauser & Fitch (2003), Lieberman argues for more focus on comparative data;
- Dunbar (2003) complains about lack of attention for social function of language;
- Komarova & Nowak (2003) identify two popular misconceptions: the view of language as a undecomposable unit, and the idea that language evolved from scratch when the human lineage diverged from the chimpanzee lineage some 5 million years ago.

If scholars within the language evolution field are critical of each other, the criticism of the field – and more broadly of the whole Darwinian approach to explaining human cognition and behaviour – from researchers outside language evolution can be withering. Linguist Noam Chomsky (quoted in Pinker & Bloom, 1990) writes:

> *"It is perfectly safe to attribute this development [of innate mental structure] to "natural selection", so long as we realize that there is no substance to this assertion, that it amounts to nothing more than a belief that there is some naturalistic explanation for these phenomena." (Chomsky, 1972, p.97).*

Chomsky (2002), 30 years later, made similar remarks. Evolutionary biologist Richard Lewontin is even less respectful:

> *"Finally, I must say that the best lesson our readers can learn is to give up the childish notion that everything that is interesting about nature can be understood. History, and evolution is a form of history, simply does not leave sufficient traces, especially when it is the forces that are at issue. Form and even behavior may leave fossil remains, but forces like natural selection do not. It might be interesting to know how cognition (whatever that is) arose and spread and changed, but we cannot know. Tough luck." (Lewontin, 1998, p.130)*

Are Chomsky, Lewontin and other critics overly pessimistic about the feasibility of thorough, scientific investigation of language origins? Perhaps not; it could be that there really is a paucity of data, and that ultimately there will be multiple scenarios of the evolutionary history of language that are coherent and consistent with the empirical facts. However, I believe these critics are premature with their assessment. Evolutionary biology has, from the days of the "modern synthesis"

(Fisher, 1930; Wright, 1931; Haldane, 1932; Dobzhansky, 1937), used two main sources of empirical evidence – genetic and comparative – and made extensive use of mathematical modelling. For language, genetic studies have only recently started to play a role (Lai, Fisher, Hurst, Vargha-Khadem & Monaco, 2001). Comparative claims have been central to language evolution research, but so far often based on surprisingly little solid empirical research (as Hauser, Chomsky & Fitch, 2002, argue). Mathematical – and computational – modelling of both the biological and cultural evolution of language has only recently started to be seriously undertaken (Grafen, 1990; Kirby, 2002b).

1.2 How to Study the Evolution of Language

Language evolution is of course not the only field where it is difficult to find empirical evidence: in cosmology, general relativity, paleontology, origins of life and many other fields researchers have struggled to find ways to test the coherence of their theories, and to test the sometimes very indirect predictions that follow from them. The solution in these fields has not been to abandon the interesting questions, but to *formalise* the theories, and to work out *testable consequences*, even if it requires many intermediate steps. For the evolution of language this requires the development of complete and formal scenarios that explain the evolution of the unique features of human language (which are testable in modern humans) from a plausible precursor state in the human lineage that is not unique in nature (and hence, open to empirical investigation through comparative research).

Only when we have precise scenarios of the evolution of language and worked out ways to test empirically the plausibility of one scenario against another, can we conclude – if that turns out to be the case – that there are too many alternative scenarios consistent with the available data. In my view, we have certainly not reached this stage yet. In this thesis I work out a number of formal requirements for theories of language evolution, and argue that existing models and theories – including models presented in this thesis – do not yet meet all requirements. The thesis is complementary to interesting work arguing that much more empirical data can and should be gathered, and reporting results from such studies (Hauser, 1996; Hauser *et al.*, 2002; Fitch & Hauser, 2004).

Of course, many other researchers have emphasised the need for scenarios of language evolution to be (i) testable (for instance, Lieberman, 1984), (ii) complete (for instance Botha, 2003; Bickerton, 2003b) and (iii) formal (for instance, Batali, 1998; Steels, 1997; Nowak *et al.*, 2002). Of those features, formalisation is perhaps most controversial, at least in the way this has been worked out in current models. Derek Bickerton, for instance, has been vocal in his criticism of the oversimplifications in mathematical and computational models (e.g. Bickerton, 2003b). There

are two responses to such criticism, formulated nicely by Cavalli-Sforza & Feldman (1981) and Batali (1998). The first emphasises the precision that comes with formal models:

> *"Our position, however, is that a mathematical theory is always more precise than a verbal one, in that it must spell out precisely the variables and parameters involved, and the relations between them. Theories couched in nonmathematical language may confound interactions and gloss over subtle differences in meaning. They avoid the charge of oversimplification at the expense of ambiguity." (Cavalli-Sforza & Feldman, 1981, p. vi).*

The second response emphasises the heuristic value of formal models, that helps the researcher to explore consequences of a set of assumptions that might me overlooked in verbal theorising:

> *"Mathematical and computational models provide a way to explore alternative accounts of the emergence of systems of communication. If the consequences of a model are consistent with expectations based on intuitions or speculation, they might obtain a small measure of support. But more interestingly (and, as it happens, more often), the consequences of a model may deviate from expectations. In working out the reasons for the differences, one can potentially develop a richer set of intuitions. Models are thus valuable to the degree that they explicitly illustrate the consequences of the set of assumptions they embody. This may be even more important than whether those assumptions are correct." (Batali, 1998, p.406).*

Both the precision and the exploration aspect of formal modelling will play a role in this thesis.

1.3 Related Approaches

The goal of this thesis is to contribute to formal, testable and complete scenarios of the evolution of human language. One can identify at least three research traditions with similar goals.

The first is the work of **Luc Steels** (since 1995) and his students and colleagues at the Free University Brussels and the Sony Computer Science Laboratory in Paris (Steels, 1995, 1998; Steels, Kaplan, McIntyre & Van Looveren, 2002; de Boer, 1999; De Jong, 2000; Kaplan, 2000; Vogt, 2000; Belpaeme, 2001; De Beule, Van Looveren & Zuidema, 2002; Oudeyer, 2003). This work is based on the conviction that very little about human language is innate, language-specific and shaped by natural selection. Rather, language – with all its complex features – is the result of the cultural negotiation of a communication system between agents with the communicative

intentions and the cognitive, perceptual and motor abilities of humans. Much emphasis is put on the biophysical constraints of "embodiment", and the spontaneous emergence of structure in "self-organisation". The methodology is described as "understanding by building" (Pfeifer & Scheier, 1999). In this work researchers try to simulate in as much detail as technically possible the emergence of the features of natural language semantics, phonology, pragmatics and syntax. The ultimate goal of this line of research is the simulation of the birth of a complete language in a population of talking robots.

The approach I will take in this thesis differs in two important ways from the research in this tradition. The first difference is methodological: I will not try to simulate reality but rather try to design simple models, that deliberately abstract out those aspects of reality that are seen as non-essential for the phenomenon under study. I believe the value of modelling is to aid understanding, and that "to substitute an ill-understood model of the world for the ill-understood world is not progress" (Boyd & Richerson, 1985, p.25). The second difference is about the role of Natural Selection, as will be emphasised several times in this thesis. I share the belief of Steels and colleagues that "cultural evolution" and "self-organisation" play a crucial role in creating the structure of languages. However, I see a complementary role for natural selection in tinkering with the parameters of self-organising processes[2].

The second tradition is the work of **Martin Nowak** (since 1999) and co-workers (Nowak & Krakauer, 1999; Nowak, Krakauer & Dress, 1999; Plotkin & Nowak, 2000; Nowak, Plotkin & Jansen, 2000; Trapa & Nowak, 2000; Nowak, Komarova & Niyogi, 2001; Komarova & Nowak, 2001; Komarova, Niyogi & Nowak, 2001; Nowak, Komarova & Niyogi, 2002; Mitchener & Nowak, 2002; Komarova & Nowak, 2003). These authors present mathematical models – some very simple, some rather complex – to clarify the major steps in the evolution of language: discrete repertoires of speech sounds, word formation, sentence formation, Universal Grammar. The similarity in ambition to the work in this thesis manifests itself even in the choice of titles, such as *Major transitions in language evolution* (Plotkin & Nowak, 2001) and *Evolutionary biology of language* (Nowak, 2000). The major differences are technical, as will become clear from the quite detailed critique of some of these models in chapters 4, 5 and 6. A recurring theme is that these models keep the representation of language very abstract, whereas I will – using more concrete representations of language and simulation models – argue that the assumed sets of strategies available for evolution are often unrealistic.

[2]De Boer (p.c.) and Oudeyer (2003) do express similar views.

The third tradition is the work of **Jim Hurford** (since 1989), **Simon Kirby** (since 1994) and **John Batali** (since 1994) and their students (Hurford, 1989; Batali, 1994; Kirby, 1994; Oliphant & Batali, 1996; Kirby & Hurford, 1997; Yamauchi, 2001; Smith, 2003b; Brighton, 2003; Smith, 2003a). These models cover a range of topics, in particular the biological evolution of lexical learning, the cultural evolution of syntax and the "learning guided evolution" of syntax (the Baldwin effect). The focus is on detailed analysis of specific simulation models, with a shift from evolutionary game-theoretic models in 1989 to mainly cultural evolution models later[3]. The models I will present differ from this tradition in that they focus much more on the role of natural selection, and on a complete scenario that includes the evolution of phonology.

1.4 Plan of the Thesis

In this thesis I will discuss theories, models and results from fields ranging from population genetics to comparative linguistics. In the interest of readability, I will avoid as much as possible the technical jargon from particular fields, limit mathematical details to a fairly basic level, and provide wherever I can a summary in words of given equations, or the intuition behind a given formalism or simulation.

In **chapter 2** I will review foundational models from evolutionary biology, to arrive at a list of formal criteria for evolutionary scenarios. These criteria, and some of the terminology introduced, will play a role in the rest of the thesis in evaluating existing work and designing new models.

In **chapter 3** I introduce a gradual scenario for the evolution of language proposed by Jackendoff (1999, 2002). Although by no means uncontroversial, this scenario is an example of the kind of complete scenarios I have in mind. I introduce a number of formalisms to characterise the various stages in the scenario, and list three transitions that need further investigation.

In **chapter 4** I study the first of these transitions, the evolution of combinatorial phonology. I will use the requirements from chapter 2 to critically evaluate existing models. I then present a new model of the evolution of this fundamental feature of speech, where speech signals are modelled as trajectories through an acoustic space. The model uses a hill-climbing heuristic to minimise confusion probabilities, and I will consider both optimal configurations and "evolutionary stable states".

Chapter 5 is very similarly structured, but considers the evolution of compositional semantics. I evaluate existing models and present a new model that is also based on a hill-climbing heuristic and uses a matrix representation for describing the mapping from meanings to signals. I show that in both the optimal configurations and the evolutionary stable states similar meanings

[3] Kirby & Hurford (1997) and Smith (2003b) combine biological and cultural evolution.

will be expressed by similar signals, and discuss the relevance for the evolution of compositionality.

Chapter 6 considers the most difficult topic, the evolution of recursive, hierarchical phrase-structure. I review some existing approaches that are based on mathematically convenient simplifications. I argue that these simplifications wrongly exclude the effects of cultural evolution. I illustrate these points with a new model of the learning and cultural evolution of phrase-structure, and consider the requirements for a model of the biological evolution of learning strategies for phrase-structure.

Finally, in **chapter 7** I discuss the implications of the models in this thesis for the debate between empiricists and nativists, and for the status of language universals. Moreover, I will argue that the formal requirements from evolutionary theory and linguistics do constrain scenarios of the evolution of language. The search for the first formal, complete, coherent and testable scenario is still open, but some steps towards such a scenario were taken.

CHAPTER 2

The evolutionary biology of language

What are the requirements for scenarios of the biological evolution of language? In this chapter I survey a number of simple but fundamental models from population genetics, evolutionary game-theory and social evolution theory. This review yields a list of required elements of evolutionary explanations in general, and of explanations for language and communication in particular.

2.1 Introduction

There are two distinct ways in which the study of evolution and the study of natural language overlap. First, they overlap in the search for an evolutionary explanation for why humans, and humans alone, are capable of acquiring and using natural languages. Second, the process of evolution in biology and the historical process of language change bear many similarities, and these parallels have played a role in the development of theories in both fields since the time of Darwin. I will throughout this thesis refer to these issues as the *biological evolution of language* (or "the language faculty") and the *cultural evolution of language(s)* respectively.

Both issues have received a great deal of attention in recent years, leading to a plethora of theories and models (Hurford *et al.*, 1998; Christiansen & Kirby, 2003a). Many proposals involve a single mechanism or factor responsible for the emergence of modern natural languages. In some cases, extensive scenarios for the evolution of language are proposed. Although this enormous body of work contains a great number of interesting ideas and findings, there are also a number methodological problems. First, it is extremely difficult to relate separate proposals to each other, because of a lack of consensus on terminology and basic assumptions. Second, it is extremely difficult to evaluate the internal consistency and empirical validity of proposed theories, because of a lack of formal rigor.

In some ways this situation is reminiscent of the state of the whole field of evolutionary biology before the establishment of theoretical population genetics by Fisher, Wright, Haldane and others in the 1920s and 30s. Their mathematical models, and the subsequent informal "modern synthesis", convinced biologists of the central role of natural selection in evolution. Confusion remained about the units of selection, but with the settling of the group selection debate by Maynard Smith (1964) and Williams (1966) a relative consensus emerged about the minimum requirements for evolutionary explanations, as well as a common vocabulary in which disagreements can be phrased. In the interdisciplinary field of language evolution, this clarity is yet lacking. In this chapter, I will review some simple mathematical models from evolutionary biology, and evaluate how they can be applied to both the biological and the cultural evolution of language.

I will start with some classical results from population genetics, about the way gene frequencies in a population change as a result of mutation and selection, and then discuss the case for viewing natural selection as optimisation, as well as the problems with this view. This optimisation view then provides a natural bridge to evolutionary game theory, where the targets of optimisation shift because the opponents in the game evolve as well. Finally, extensions to social evolution models that deal with kin selection, will lead us to the issue of levels of selection, and

clarify the relation of cultural evolution models – with the dynamics happening at the level of cultural replicators – to evolutionary biology generally.

2.2 Adaptation for Language

When chimpanzees, our closest living relatives, are taught human language, they acquire several hundreds of signals (Gardner & Gardner, 1969; Savage-Rumbaugh *et al.*, 1986). They fail, however, to produce speech sounds themselves, to acquire the many tens of thousands of words in natural languages, and to grasp the use of even the most basic rules of grammar (Terrace, 1979). Human infants, in contrast, acquire their native language rapidly. They produce speech sounds and comprehend simple words before the age of 1, produce their first words soon after their first birthday and the first grammatical constructions before their second birthday (Tomasello & Bates, 2001).

Why? Clearly there is something special about humans that makes them extra-ordinarily apt to acquire and use natural languages. Among other things, the anatomy of the vocal tract, the control mechanism in the brain for complex articulation and the cognitive ability to analyse and produce hierarchically structured sentences appear to be qualitatively different in humans than in other apes. But not only humans are special; there is also something special about natural languages that makes them extra-ordinarily apt to be acquired and used by humans.

How did this tight fit come about? One possibility is that the human capacity for language has emerged purely as a side-effect of the many changes in anatomy and cognition that occurred in the hominid lineage. The tight fit itself, in such a scenario, doesn't need to be accidental, because a cultural evolution scenario predicts that language will adapt to the peculiar biophysical and cognitive features of humans that themselves have evolved for other reasons.

Although this possibility cannot be dismissed, from a biological point of view it does not appear very likely. Humans spend around 3 hours a day or over 20% of their awake time talking (Dunbar, 1998, and references therein), verbal abilities play a significant role in social status and, it seems, in both the reproductive success of individuals and the success of our species as a whole. Such a salient characteristic of any organism would require a Darwinian, evolutionary explanation. Hence, although the side-effect option is a possibility, it can only be the conclusion of an elaborate investigation, and not serve as null hypothesis. In chapter 6 I will argue that although language as a whole might be considered a biological adaptation, many specifics about language (language universals) are better understood as the outcome of cultural evolution. In this view, the complex results of cultural evolution and social learning have had indirect consequences for biological evolution.

If we want to investigate specific hypotheses on adaptations for language, what form should such hypotheses take? The early formal models in population genetics are a useful starting point. But first, it should be clear that any statement about biological evolution is a statement about how genes mutate and spread in a population through random drift and selection. That statement in no way reflects the form of genetic determinism or naivety about "language genes" that have made some evolutionary linguists wary to talk about genes at all. But if properties of language are to be explained by some biological endowment, which in turn is to be explained as an adaptation for language, then we need to be explicit and postulate a series of altered genes that influence the ability for language. Such genes can have many additional non-linguistic effects (an illustrative example is the recently discovered FOXP2 gene, that, when mutated, causes a range of problems in language processing as well as in sequencing orofacial movements, Lai *et al.* 2001). We can phrase this requirement[1] as follows:

Criterion 1 (Heritability) *Evolutionary explanations for the origins of a trait need to postulate genetic changes required for that trait.*

2.3 Evolution as Gene Frequency Change

A formal model of evolution as gene frequency change can be built-up in the following way. Consider first that in humans, as in almost all multicellular organisms, every individual inherits two sets of genes, one from the father and one from the mother. If there is to be any change, we need to consider at least two different variants, alleles, for each gene locus, and monitor the increase in frequency of one allele at the expense of the other. In figure 2.1 the Mendelian model of inheritance of two alleles – A and a at a single locus – is depicted. Adults (top row) have a genome that is of any of the three possible types AA, Aa or aa (Aa and aA are equivalent). These adults produce sperm and egg-cells (second row) with just a single copy of the gene under consideration. In sexual reproduction, a sperm-cell and an egg-cell fuse, and grow out to a new individual (third row). Evolution, in this simple scheme, concerns the change in frequencies of the genotypes AA, Aa or aa, or the change in frequencies of the alleles A and a.

The Hardy-Weinberg model (developed independently by British mathematician Godfrey Harold Hardy, 1908 and German physician Wilhelm Weinberg, 1908; see Crow, 1999) describes the gene frequencies if there is no mutation or selection. Consider the frequencies of the three

[1] Of course, one can sensibly study the evolution of traits for which the genetic component has not been identified. The point here is to emphasise that biological evolution implies genetic changes. The "requirements" in this chapter concern the ultimate evolutionary explanation for a trait; of course, not every evolutionary model study will be able to meet all requirements, and neither will the studies presented later in this thesis.

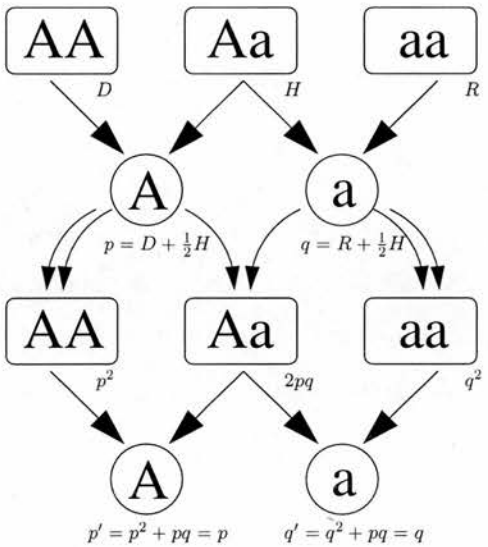


Figure 2.1: Mendel's model of inheritance, and the Hardy-Weinberg model of allele and genome frequencies under Mendelian inheritance with no selection nor drift.

genotypes (top row) at any particular point in time, and call these frequencies D, H and R. The frequencies of the alleles A and a in the sperm and egg-cells are simply:

$$\begin{aligned} \textit{frequency of } A: \ p &= D + \frac{1}{2}H \\ \textit{frequency of } a: \ q &= R + \frac{1}{2}H, \end{aligned} \tag{2.1}$$

because individuals with genotype AA or aa will always pass on an A or a respectively to their sperm and egg-cells, but individuals with genotype Aa only half of the time.

Under a number simplifying assumptions (including random mating and meiosis, an infinite population and no sex differences at the relevant locus), the frequencies of the three genotypes in the offspring are simply $D' = p^2$, $H' = 2pq$ and $R' = q^2$, because you need two A's or a's to make an AA or aa respectively, and you need an A from either the father or the mother and an a from the other parent to make an Aa. When this offspring then starts producing sperm- and egg-cells, the frequencies of the alleles A and a are:

$$\begin{aligned} \textit{new frequency of } A: \ p' &= D' + \frac{1}{2}H' = p^2 + pq \\ \textit{new frequency of } a: \ q' &= R' + \frac{1}{2}H' = q^2 + pq. \end{aligned} \tag{2.2}$$

Hardy and Weinberg's simple but fundamental observation is that because $p + q = 1$ (the total frequency of all alleles must be 1, and thus $q = 1 - p$), it follows that p and q are constant under

this model of inheritance:

$$p' = p^2 + pq = p^2 + p(1-p) = p^2 + p - p^2 = p. \tag{2.3}$$

This result shows that under Mendelian inheritance existing variation in gene frequencies is maintained. This is in contrast with "blending inheritance" (the assumed model of inheritance before the rediscovery of Mendel's laws around 1900), where a child's trait values are the average of the parents' and variation quickly dissipates over time. The result played a crucial role in reconciling Mendelian genetics with Darwinian evolutionary theory, because it showed that under reasonably low mutation rates enough variation can be built up for natural selection to operate (Fisher, 1930, chapter 1).

The Hardy-Weinberg model can be extended in a straightforward manner to include the effects of selection. Natural selection, in Darwin's theory, is the consequence of differences in survival rates to the age of reproduction and the differences in reproductive success. These effects can be summarised with a fitness coefficient for each of the possible genotypes, which gives the expected number of offspring. A high coefficient w_{AA} means that individuals of genotype AA live long and reproduce successfully, such that their genes are well represented in the next generation. In terms of the equations, this just requires weighting the contributions of parents of each genotype with the relevant fitness coefficient:

$$p' = \frac{p^2 w_{AA} + pq w_{Aa}}{\overline{w}}, \tag{2.4}$$

where $\overline{w}$ is the average fitness and given by:

$$\overline{w} = p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa} \tag{2.5}$$

(this term is needed to account for changes in population size due to reproduction and selection).

Equation (2.4) gives us a first handle on the requirements for evolutionary innovation, and, hence, evolutionary explanations. First of all, natural selection operates on genotypic and phenotypic variation. Second, natural selection favours fitter genes and individuals over less fit ones. Both the variation and the fitness differences need to be made explicit:

Criterion 2 (Strategy set) *Evolutionary explanations need to postulate a set of possible genotypes and phenotypes, as well as the mutations that can move an organism from one genotype–phenotype to another.*

Criterion 3 (Payoff function) *Evolutionary explanations need to postulate a function that relates the possible genotypes–phenotypes in a given environment (that may include other evolving individuals) to fitness.*

If we are interested in a specific biological innovation – that is, a mutation – that was relevant for learning or using language, we need to consider the situations before and after that mutation. In the simplest case, a is the preexisting gene that is initially shared by the whole population, and A is the mutated version of a that has arisen in a single individual. Hence, initially $q \approx 1$ and $p \approx 0$. If A is to play a role in an evolutionary scenario, we need to establish that allele A did start to spread in the population (as sketched in figure 2.2); in other words, that p increases. We can formulate this requirement as follows:

Criterion 4 (Invasibility) *Innovations in an evolutionary scenario need to be able to invade a population; that is, an innovation should spread in a population where it is extremely rare.*

If we know all fitness coefficients, it is straightforward to work out what happens to the frequency of the new mutation. As it turns out A will spread if $w_{Aa} > w_{aa}$, and it will get fixed ($p = 1$) if $w_{AA} > w_{Aa}$. In other words, the fitness of the new gene must be greater than that of the old one, and the new gene must, to some extent, be *dominant* over the old one such that its effects are noticed in individuals that inherit copies of both genes from each of the parents. In fact, the difference in fitness between the two variants must be significant, at least large enough for the new gene not to get lost by chance fluctuations (Fisher, 1922) and to get established after a reasonable number of generations (Haldane, 1932). Note that these results depend on some strong assumptions, including an infinite population with randomly interacting individuals. In small populations with non-random interactions different dynamics can occur.

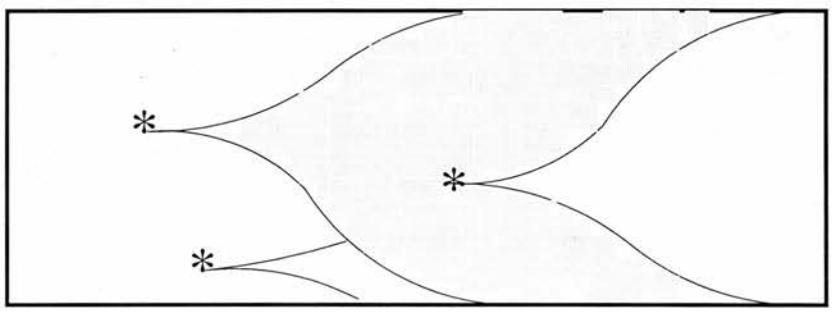


Figure 2.2: The spread of new genes in a population

2.4 Evolution as Optimisation

Since Darwin (1859), the notion of "adaptation" has played a major role in evolutionary thinking. His work offered a coherent framework to study the traits of organisms in terms of their *function* for survival and reproduction. Even before the mechanisms of genetic inheritance were unravelled, Darwin thus transformed biology from a descriptive to an explanatory science. In the early 1920s the "founding fathers" of population genetics – Fisher, Wright and Haldane – worked out what happens to a single new gene when it appears in a population. But do the dynamics described by equation (2.4) constitute "adaptation"? In other words, does the predicted change in gene frequencies also mean the population will get better adapted to its environment, i.e. improve its average fitness?

Both Fisher and Wright set out to work out a more general result. I will discuss Fisher's "fundamental theorem of natural selection" (Fisher, 1930) in section 2.8. Here I will follow Wright's analysis of the average fitness in a population, in particular Roughgarden's (1979) version of these equations. Most mathematical details are in appendix A, but it is useful to look at a couple of Wright's equations. First, it is convenient to look at the *change* in the frequency p at every timestep. This is, using equation (2.4), given by:

$$\begin{aligned} \Delta p &= p' - p \\ &= \frac{p^2 w_{AA} + pq w_{Aa}}{\overline{w}} - p \end{aligned} \tag{2.6}$$

This equation can, with a bit of algebra (see equations (A.4) and (A.5) in appendix A), be rewritten as follows:

$$\Delta p = \frac{pq}{\overline{w}} \left(p \left(w_{AA} - w_{Aa} \right) - q \left(w_{aa} - w_{Aa} \right) \right) \tag{2.7}$$

This equation tells us nothing new; it is essentially equation (2.4) in a different form. However, the new form will prove useful when we have worked out the next equation. We are interested in what happens to the average fitness when the frequency (p) of the innovation changes. Mathematically, that question directly translates into the derivative of $\overline{w}$ with respect to p. The expression for average fitness is given in equation (2.5). Its derivative, if we assume the fitness coefficients are independent of p and q (that is, no frequency-dependence) turns out to be (as is worked out in equation (A.2) and (A.3) of appendix A):

$$\frac{\mathrm{d}\overline{w}}{\mathrm{d}p} = 2 \left(p \left(w_{AA} - w_{Aa} \right) - q \left(w_{aa} - w_{Aa} \right) \right) \tag{2.8}$$

When we note that equations (2.7) and (2.8) are very similar, it is clear that we can replace a large part of (2.7) with half of (2.8), and get:

$$\Delta p = \frac{pq}{\overline{w}} \left(\frac{1}{2}\right) \frac{\mathrm{d}\overline{w}}{\mathrm{d}p}. \tag{2.9}$$

This is a fundamental result for evolutionary biology. The equation says that the change in the frequency of a new gene, will be *in the direction* of the derivative of fitness with respect to that gene's frequency. That means that only if the average fitness increases with increasing p, will the new gene spread. Moreover, the spread will be fastest at intermediate frequencies (high variance) and low average fitness. In other words, evolution – under the assumption mentioned – will act to optimise the average fitness in the population: it will lead to adaptation.

However, the mathematical derivation of this intuitive result also tells us about its limitations. First of all, evolution is shortsighted. We saw a simple example at the end of the previous section: if $w_{Aa} < w_{aa}$ (there is "heterozygous disadvantage"), then the new allele A will not spread in the population, even though at fixation it might improve the mean fitness in the population. Second, evolution needs (heritable) variation. If $pq = 0$, nothing will change. Thirdly, the equation is only valid if the fitness coefficients are *independent* of p and q. That is, whatever the traits are that allele A influences, the usefulness of the innovation should not depend on how many others in the population share it. This condition is obviously violated in the evolution of communication, because the usefulness of a signal will always depend on the presence of others that can perceive and understand it. Fourthly, the original Hardy-Weinberg model brought quite a lot of assumptions, including the independence of the single locus we looked at from other loci, random mating, discrete generations and infinite populations. Some of the consequences of relaxing these and the frequency independence assumptions will be evaluated in the next section.

Finally, as Fisher (1930) emphasised, these calculations deal only with the direct effects of natural selection. They predict the direction of change, but it is unwarranted to conclude that the average fitness in a population will increase. Environmental conditions might have changed in the mean time and, even if the environment is constant, all individuals in the population are better adapted to it such that competition is fiercer. These effects – not modelled by Wright and Fisher's equations – were collectively labelled "deterioration of the environment" by Fisher.

In addition to these quantitative results, Wright made a much more qualitative contribution relating evolution and optimisation. In a paper without any mathematics (Wright, 1932) he introduced an extremely influential metaphor: the **adaptive landscape**. The adaptive landscape is a landscape of 3 or more dimensions, with the plane (or hyperplane) representing the space of

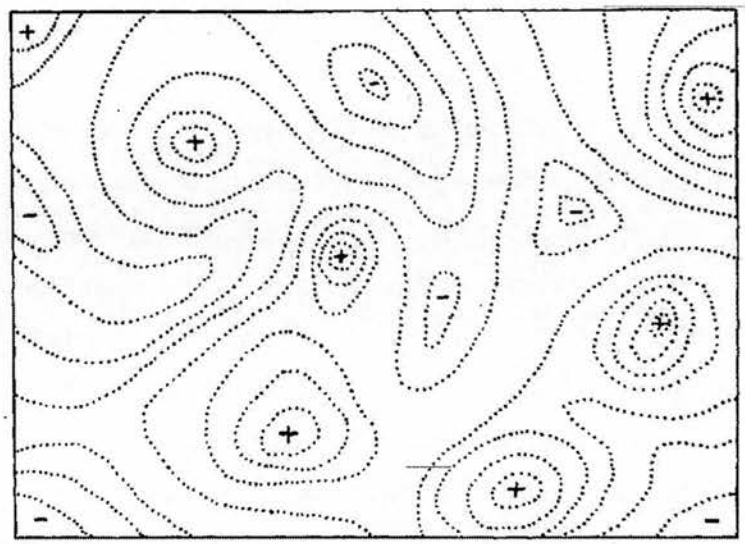
possible genotypes, and the height of every point representing fitness (see figure 2.3). On such a landscape, a population is a collection of points. Mutations correspond to steps in the landscape; selection corresponds to the selective removal of individuals that are lower down. The process of evolution involves the population to climb up-hill, following a local gradient to a local peak.

I will discuss some problems with the concept below. However, the adaptive landscape representation in this form does illustrate Darwin's (1859) insight that for a process of continuing evolution, we need a path of ever increasing fitness from the hypothesised initial point in genotype space to the end result. (In finite populations, stochastic drift can bridge fitness barriers in the adaptive landscape, but only if they are relatively shallow.) For complex traits, such as language, it seems reasonable to postulate a series of many genetic changes. Wright's metaphor highlights the fact that each of these changes needs to confer an adaptive advantage:

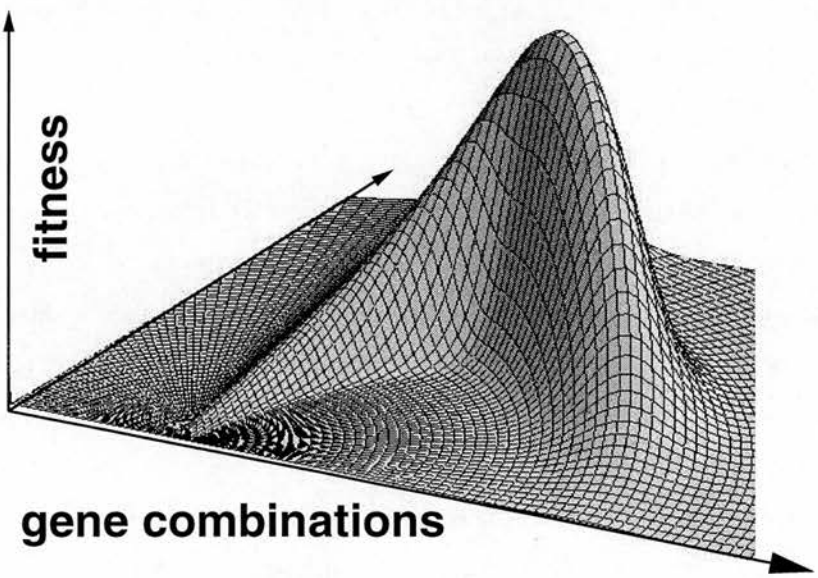
Criterion 5 (Fit intermediates) *Explanations for complex traits, that involve a series of genetic changes, need to show a path of fit intermediates, from the hypothesised initial state to the desired end state.*

This requirement is important, but it might not be as problematic as it looks at first sight. First, although evolution will generally lead uphill, there is some room for random processes as well. Wright used the adaptive landscape metaphor to explain the effects of increases or decreases of the rate of mutation and the strength of selection. He also discussed at some length the effects of small population sizes, where inbreeding will lead to the non-selective process of genetic drift: random deviations from the locally optimal genotype due to accumulation of mutations and a lack of variation for selection to operate on. Wright's shifting balance theory (or at least one version of it) argues that the additional variation inherent in subdivided and inbreeding populations could help the population as a whole bridge fitness barriers. Although the shifting balance theory has little empirical support (Coyne, Barton & Turelli, 2000), the basic idea that, under some conditions, genetic drift could help bridge a fitness barrier remains.

Second, one of the basic tenets of evolutionary biology is that all life originates from the same source. If that is true, all complex traits of all organisms are connected through paths of fit intermediates. Thus, if we wonder if there is a path on the adaptive landscape through which humans could evolve wings, the answer must be yes. Humans, bats and birds have a common ancestor, so there must be at least one series of environments (including other species) that would yield a path that leads from humans back to the common ancestor with bats, and again forward to modern bats (ignoring some difficulties such as frequency-dependent fitness).



(a) Wright's graph of the adaptive landscape



(b) A computer-generated 3d adaptive landscape

Figure 2.3: The adaptive landscape of fitness as a function of genotype. The graphs illustrate hypothetical examples in which two genes have a continuous range of effects. Real organisms have, in contrast, a discrete set of possible genotypes involving many more than two genes. Thus, mutations can take them in very many directions. This high dimensionality makes it more likely that there is some path uphill to the "adaptive peak" (see Provine (1986), chapter 9). (a) is a graph from Wright (1932). The original caption is: "Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness." (b) is taken from Barton & Zuidema (2003).

Third, intuitions about getting stuck in local peaks based on the three-dimensional representation as in figure 2.3 must be treated with care. There are, in fact, a great number of problems with the concept (Provine, 1986, in his biography of Wright, gives a thoughtful critique). First of all, as Wright indicated, an actual genome consists of many (tens of) thousands of genes. Hence, the adaptive landscape has tens of thousands of dimensions, rather than just 3. That makes a big difference, because whereas local peaks seem extremely likely in 3 dimensions, they are in fact increasingly less likely with more and more dimensions. But, perhaps more importantly, the genotype space in Wright's graph is continuous, whereas the genotypes of actual organisms are discrete. Wright's landscapes, as drawn here, can in fact never be constructed for a real example.

Wright and others have looked at other versions of the adaptive landscape that are, in contrast, rigorously defined. One approach is to choose the gene frequencies and population average fitness as axes. A population, in this representation, is then a single point in the landscape. The advantage of this representation is that it ties in nicely with the mathematical model of equation (2.9). However, the disadvantage is that in such a landscape one cannot visualise the effects of selection, mutation, genetic drift and subdivision of the population, which was the whole point of introducing the metaphor.

Alternatively, one can choose to use phenotypic, continuous traits against individual fitness as the axes of the landscape. The disadvantage of this approach is that mutations, which define what a genotype's "neighbours" are, are of course defined genotypically. Therefore, the random variation that builds up by mutation, will not generally be centred around a single population mean in phenotypic space. In cases where very little is known about the genetics anyway, such as language, that might not really matter, but, as we will see, there the landscape cannot be constructed anyway because of frequency dependence.

Nevertheless, the view of evolution as optimisation yields a powerful approach for deriving predictions about an evolving system, or for understanding an evolved system as adapted for a specific purpose. Parker & Maynard Smith (1990) present a methodology for evolutionary reasoning based on this view which they call "optimality theory"[2]. They first emphasise that every evolutionary study must start with identifying a clear biological question. Step 2 is to identify a set of strategies that are available for evolution to choose from. Step 3 is to identify a pay-off function, which evolution is supposed to optimise, and to show that the observed biological phenomenon tends towards the optimum. Step 4 is to relate pay-off, which is an indirect measure for fitness, to actual fitness. Finally, step 5 is to derive predictions and test them empirically.

[2]Parker & Maynard Smith's (1990) Optimality Theory is completely unrelated to Optimality Theory (Prince & Smolensky, 2004) in linguistics.

This scheme provides a coherent framework for thinking about the evolution of language, and it is essentially the approach I have taken in this chapter and the rest of the thesis, although I have and will put some extra emphasis on specific implications of the approach relevant for language evolution. Note however, that the mathematical models discussed so-far concerned changes in gene frequencies, whereas Optimality Theory and language evolution research are concerned with phenotypic traits that typically involve many, often unknown genes. I will first discuss some limitations of the optimality view that apply even when we look at traits controlled by a single gene, and then discuss the more difficult issue of going from single-gene models to the evolution of complex phenotypic, traits such as language.

2.5 Limits to Optimality

> *"Natural selection tends only to make each organic being as perfect as, or slightly more perfect than, the other inhabitants of the same country with which it comes into competition. And we see that this is the standard of perfection attained under nature" (Darwin, 1872, p 163; quoted in Provine 1986, p209).*

As Darwin was well aware, the fact that evolution can be understood as optimisation does not imply that the features of organisms are optimal or perfectly adapted to their environment. The most obvious evidence for the existence of limits to optimality, are the many examples of indigenous species that are rapidly driven to extinction after humans introduced a foreign competing species. There is a whole tradition of listing the limitations of natural selection (e.g. Dawkins, 1982; Barton & Partridge, 2000). These can be roughly classified in four classes: (i) biophysical and genetic constraints, (ii) the speed of evolution, (iii) mutational load and (iv) fluctuating fitness.

With regard to **biophysical constraints**, it is clear that all of the complexities of biological organisms need to grow out of a single cell. Throughout its development, an organism needs to maintain its metabolism, to selectively take up chemicals from its environment and to autonomously build-up all of its complex features. That process of biological pattern formation is constrained by what is possible at all with the materials available in a biotic environment, by what can be coded for by genes, and by which possibilities are reachable for evolution. It is obvious that these constraints are at work, given for instance the limitations in speed of both a prey and a predator trying to outrun each other. It is also obvious, however, that these limitations have not prevented evolution from building exquisitely complex and well-adapted organs such as, for instance, the human ear (see chapter 3, section 3).

Population and molecular genetics make some specific predictions on **genetic constraints**. Natural selection can often not optimise all different phenotypic traits independently from each other, because of the following features of genes:

- A single gene typically has an effect on many different phenotypic traits (pleiotropy);
- The effect of a gene on a trait depends on the presence or absence of other genes (epistasis);
- Genes are physically linked to each other in a chromosome (linkage).

The little that is known about human genetics relevant for language (e.g. Lai *et al.*, 2001) suggests, unsurprisingly, that all these general observations hold for language as well. The general observation have played a role in a debate about whether or not the Baldwin effect – where initial learnt traits are "assimilated" by genetic evolution – is likely to have played a role in the evolution of complex language (Hinton & Nowlan, 1987; Briscoe, 2000b; Yamauchi, 2001; Briscoe, 2003) Nevertheless, it seems too little is known about human genetics to inform specific models of the evolution of language, so they will not play a role in the rest of this thesis.

Most of these biophysical and genetic constraints are reflected in the choice of the strategy set, which contains all strategies/trait values that are available to evolution, and excludes those that cannot be instantiated. The physical linkage between genes, however, is – in the long term – not one of these hard constraints on what can evolve, because recombination will eventually break the linkage such that one gene can occur without the other. Linkage does constrain how fast things can evolve, which is also crucial for the course of evolution.

More generally, the **speed of evolution** is constrained by the available genetic variation at every step (including effects from linkage) and the strength of selection. Considerations about evolutionary time should be included in evolutionary explanations:

Criterion 6 (Sufficient time) *Evolutionary explanations need to establish that there has been enough time for favourable alleles to get established in the population.*

Evolution needs variation to operate on, and mutation is the source of this variation. However, because mutation is indiscriminate and random, it will also constantly create individuals that are worse than average, or even unviable. This is called **mutational load**. In the adaptive landscape metaphor, whereas selection will push a population to the top of an adaptive peak, mutation will pull the population down-hill. The dynamic equilibrium is called *mutation–selection balance.* For specific cases, such as the evolution of RNA molecules, the constraints on optimisation posed by mutational load can be worked out. For the case of language, again too little is known of its genetic basis to derive any specific limitations. However, since a series of formal models of the cultural transmission of language have been proposed (Nowak *et al.*, 2001; Komarova *et al.*, 2001;

Mitchener & Nowak, 2002) that are based on the concept of mutational load, it is worth looking in a bit more detail at how this concept has been formalised.

Eigen (1971) and colleagues generalised the Fisher-Wright equations for evolution with mutation and selection at a single locus, to dynamics with an arbitrary number of loci. Using notation loosely based on Maynard Smith & Szathmáry (1995) and Nowak *et al.* (2001), we can write Eigen's equation as follows:

$$\Delta x_i = \sum_{j=1}^{M} (x_j w_j \mathbf{Q}_{ji}) - \overline{w} x_i, \tag{2.10}$$

where i and j are indices for all the M distinct possible genotypes. Δx_i stands for the changes of the frequencies of all genotypes i (hence, the expression (2.10) defines a system of equations, all the same form and one for each possible i). x_i is the frequency of genotype i and w_i its fitness. $\mathbf{Q}_{ji}$ is the probability that a given child will have genotype i if her parent has genotype j. Hence, $\mathbf{Q}$ is an extremely large matrix of size $M \times M$ that describes the effects of mutation. Finally, $\overline{w}$ is the average fitness in the population; the last term ensures that the effects of selection are relative to the population average fitness.

Eigen looked at a very specific choice of parameters. Suppose that there is a single genotype with a high fitness, and all other genotypes have the same, low fitness. That is, the adaptive landscape is flat, except for a single high peak. Now suppose there is a constant probability μ of mutation per gene, and no cross-over. The probability q that an individual (here: an RNA-molecule) when it reproduces produces identical offspring is now:

$$q = (1 - \mu)^l, \tag{2.11}$$

where l is the genome length. q is called the "copying fidelity". With a bit of algebra one can work out where the mutation–selection balance is for different mutation probabilities, and thus different copying fidelities. Eigen's exciting result is that there is a precise value of q where the mutation–selection balance suddenly drops to vanishingly small quantities of each possible genotype. That is, if the mutation probability is above a threshold value – the *error threshold* – selection ceases to play any role, and individuals have essentially random genotypes:

Criterion 7 (Mutational load) *Evolutionary explanations need to postulate a mutation rate high enough to generate the variation needed, but low enough to not suffer from an extreme mutational load (to cross the error threshold).*

A final category of limits on optimality comes from **fluctuating fitness**, that is, from the fact that the fitness regime of organisms is constantly changing. First of all, there are temporal fluctuations in the environmental conditions on many different timescales, both regular and irregular: from the day and night cycle to climate changes. Similarly, there are geographic differences, such that migrating organisms might find themselves in very different habitats. Organisms adapted to one set of conditions, are not necessarily adapted to other conditions. A language that evolved for communication between hunter-gatherers on the savannah, is not necessarily adaptive in a modern city environment.

But perhaps more interesting is the situation where the fitness regime of a particular species changes due to evolutionary changes of the species itself (**frequency dependent selection**) or of any of the other species it interacts with (**co-evolution**). The evolution of language and communication is frequency-dependent, because linguistic innovations are unlikely to pay off if there is no one to talk to. The fitness coefficients in language evolution are therefore not constants, as in equation (2.8), but will depend on the frequencies of the different alleles in the population. Evolutionary game theory is the general framework for addressing frequency-dependent selection, and will be discussed in the next section. Because natural languages are transmitted culturally, there can also be a process of cultural evolution, such that we can perhaps sensibly speak about the *coevolution of language and the brain* (Deacon, 1997; similar ideas were explored earlier in e.g. Christiansen, 1994; Kirby, 1994). This will be explored a bit further in section 2.10 in general terms, and with a specific model of the learning and evolution of grammar in chapter 6.

A related phenomenon is **sexual selection**, where selection is not on the ability to survive to reproductive age or the ability to reproduce per se, but on the ability to beat rivals of the same sex in the competition for a mate, or on the ability to persuade potential sexual partners to choose one as a mate (Darwin, 1859, p.94). Here, the fitness of a given genotype (defining e.g. a male trait) is not fixed, but also dependent on the frequency of all the possible genotypes (regulating e.g. female preferences) in the population. Exotic, maladaptive traits that are due to sexual selection, such as the ornate peacock-tale or the violent and sometimes lethal *love darts* in hermaphrodite snails, are nice examples of the suboptimal traits that can result from frequency dependent selection. In the evolution of speech, sexual selection seems to have played a role in shaping the secondary sexual traits, such as the lower pitch in human male voices, which results from larger larynx and vocal folds, and a change in formant frequencies at puberty, which makes males appear larger and results from a second descent of the larynx. More controversial are ideas about the role of sexual selection in the evolution of the first descent of the larynx (that happens in both males and females

in the first few months after birth, Lieberman, 1984; Hauser & Fitch, 2003), and in the evolution of complex syntax (Pinker & Bloom, 1990).

2.6 Phenotypic Evolution

We have seen that evolution can be understood as a process of optimisation, but under a range of constraints and with continuously shifting targets. The constraints and trade-offs are all crucial elements of adaptive explanations. In fact, without such constraints, the notion of "adaptation" would be meaningless: without constraints and trade-offs, only almighty beings would exist. The more precise we can be about constraints and trade-offs, including about genetic details, the more convincing demonstrations of optimality within these constraints are as evolutionary explanations. However, even without a complete understanding of the genetic constraints, we can make progress in understanding evolution at the phenotypic level, by incorporating likely constraints in formal models and deriving testable predictions.

As an example of the structure of such optimality arguments, consider the evolution of hearing and suppose that it can be described with a single variable: the threshold value θ for signal detection. Presumably, the benefit is maximal when this θ approaches 0 (assuming the brain can select and process the information it needs), and the benefit approaches 0 when θ is infinitely large. The cost of an infinitely small θ is infinitely big, however, because biophysical constraints dictate that infinitely small θ requires infinitely large ears. With very large θ we could do away with ears all together and have a cost approaching 0. When we subtract the cost from the benefit, we get the payoff function. If the cost and benefit function adequately describe the selection pressures and constraints, we expect the evolutionary dynamics to lead to the optimum of the payoff function, shown qualitatively in figure 2.4. Now, if we could find a combination of benefit and cost functions, and empirical observations of θ in nature that match the predicted optimum, that would constitute strong evidence for either the hypothesis that θ evolved for the function described by the payoff function, or – if we are already confident of the adaptive function – that the hypothesised constraints, described by the cost function, were the right ones.

Can we make a similar analysis of the evolution of key features of natural language? That is, can we identify the payoff function and its optimum under relevant constraints and show that natural language corresponds to that optimum? Unfortunately, we know relatively little about the biophysical and genetic constraints, the relevant mutations in the evolution of language and the neural implementation of our linguistic abilities. It is therefore difficult to make precise what strategy set was available for evolution. The best examples of trade-offs in language are probably in the physical properties of speech. Liljencrants and Lindblom's (1972) demonstration that the

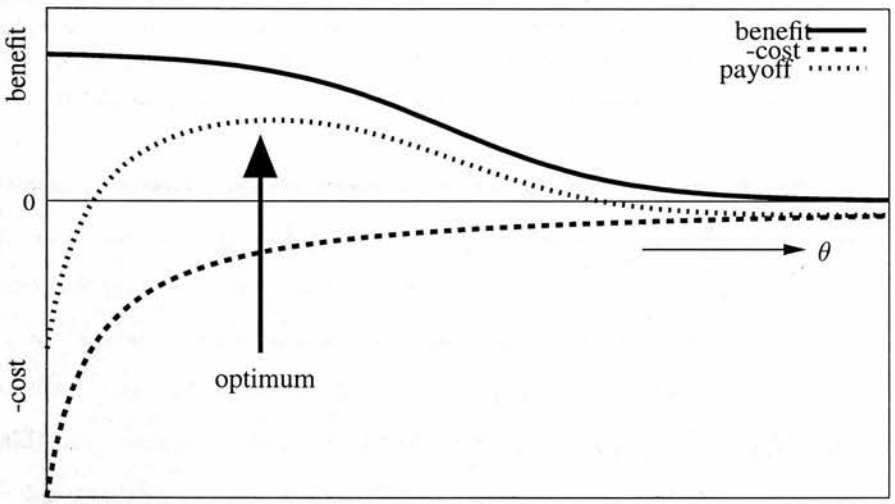


Figure 2.4: Evolutionary optimisation under biophysical constraints. The graph sketches the benefits (top curve) and costs (bottom curve) for a continuous range of detection thresholds θ (x-axis) in the evolution of hearing. An extremely low threshold (left end) is very useful, but also very costly; an extremely high threshold (right end) is very cheap, but not of much use. The optimum of the payoff function (middle curve) is therefore at an intermediate value of θ.

vowel systems in human language appear to be optimised for reliable recognition under noisy conditions and under constraints on perception and articulation, is suggestive (see chapter 4). Lieberman (1984) has argued that the human larynx has descended deeper down the throat in order to allow more flexibility of the articulatory organs. This allows us to make many different speech sounds, at the expense of an increased propensity to choke. Although controversial (Hauser & Fitch, 2003), this theory on the evolution of language does illustrate the role of evolutionary trade-offs that result from the physiological constraints in speech production.

For other components of human language, such as its semantics or syntax, it is extremely difficult to derive biophysical constraints. What sort of grammars can or cannot be encoded by genes and implemented in neuronal tissue? The only solid results relevant to this question, suggest that quite a variety of networks of interacting cells are *Turing equivalent*. That is, they can – if sufficiently large, given sufficient time and properly initialised and interpreted – compute any computable function (Siegelmann & Sontag, 1991; Wolfram, 2002). This is not to say that any grammar can be easily encoded by genes or acquired by a neural net; but without better models of the neural implementation of language, we cannot start to make sensible assumptions about the actual architectural constraints on natural language syntax that were at work during human evolution. This is how I interpret Chomsky's well-known reservations about the feasibility of scientific explanations of the evolution of language, such as expressed in this famous quote:

> *"We know very little about what happens when 10^{10} neurons are crammed into something the size of a basketball, with further conditions imposed by the specific manner in which this system developed over time. It would be a serious error to suppose that all properties, or the interesting properties of the structures that evolved, can be 'explained' in terms of natural selection." (Chomsky, 1975, p.59).*

However, it would be overly pessimistic to conclude – as Chomsky seems to do – that we can therefore not say anything sensible about how language evolved. There are two categories of constraints in language evolution that can be made precise. First of all, we have good "mentalist" models of syntax that describe its fundamental computational properties, and the **computational constraints** that any implementation will face. For instance, we know there exist constructions in natural languages that cannot be modelled by weaker formalisms (in terms of the extended Chomsky Hierarchy) than (mildly) context-sensitive rewriting grammars (Joshi *et al.*, 1991); we know that the whole class of context-sensitive rewriting grammars is not *identifiable in the limit* from positive samples alone (Gold, 1967); and we know that grammars of that type have a worst-case time-complexity of $O(n^5)$ in parsing (Barton & Berwick, 1987). Such computational constraints on representation, learning and processing, and the formalisms they are expressed in, allow us to at least make a start with testing the internal consistency of an evolutionary scenario, and with formulating a sensible strategy set for evolution. Formalisms that we can use to describe key features of human language and to derive computational constraints are discussed in the next chapter (3).

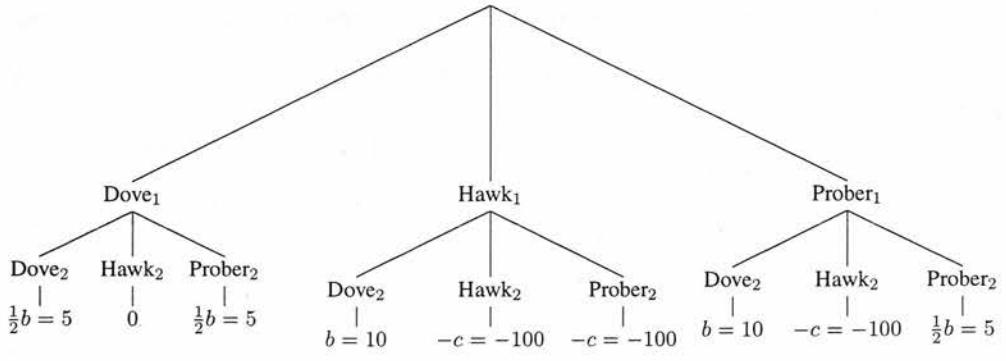
Second, there are constraints that follow from the **social, communicative function** of language. Humans use natural language to communicate with others, on the average for many hours a day per person. This requires a shared code, such that both speakers and hearers understand the meanings of utterances. Moreover, it requires the willingness of the speaker to give away information and, at least in general, to be truthful, as well as a willingness from the hearer to listen and interpret the message received. These issues can be addressed in the framework of evolutionary game theory, which will be discussed next.

2.7 Evolutionary Game Theory

The evolutionary history of human language can be viewed as a process of phenotypic optimisation, under (largely unknown) biophysical and cognitive constraints that determined which communication systems were possible at all, and in a social–communicative context that determined which systems were better than others, but that continuously shifted the evolutionary targets because the frequency of a linguistic trait in the population influences its usefulness.

The formal framework to describe the consequences of multiple agents optimising their own payoff in a social context is the **Theory of Games**. Game theory conceptualises the interaction between agents, the "players", as a game where all players choose from a set of available strategies. Crucially, the outcome of a game for each player, its payoff, depends on the strategies of other players. Unlike the example in figure 2.4, where payoff is a function of the player's own strategy alone (the trait value, θ), in game theory the payoff is a function of both the player's strategy and the strategies played by other players.

The following example is derived from Maynard Smith & Price (1973). Imagine a conflict between two birds competing for a single food source, each with the choice between three strategies: "dove" (retreat immediately if the other player is aggressive), "hawk" (always be aggressive) and "prober" (start off aggressive, but share the food source peacefully if the other player does not give up, but does not escalate either, and continue aggressively if the other player does give up). If the value of the food source is $b = 10$, and the expected cost of an escalated fight $c = 100$, the possible payoffs for player 1, given her and player 2's decisions, are given in figure 2.5(a). For 2 players and a small number of discrete strategies, this can be conveniently summarised with a *payoff matrix*, as in figure 2.5(b).



(a) extensive representation

	player 2's strategy		
player 1's strategy ↓	Dove	Hawk	Prober
Dove	5	0	5
Hawk	10	-100	-100
Prober	10	-100	5

(b) payoff matrix player 1

Figure 2.5: Extensive and matrix representations of games

We can postulate a decision mechanism for each player, and study how the outcome of the game changes with players adapting their strategies based on what the other players do. The dynamics of such games, with all players making their own decisions, are often extremely difficult to describe. Often, however, it is possible to derive the conditions under which a game is stable. In non-cooperative game-theory – where "selfish" players each try to optimise their own payoff – the crucial concept is that of a **Nash equilibrium** (Nash, 1950)[3]. This equilibrium is defined as the situation where no player can increase her payoff by unilaterally changing her strategy. Thus, for any n-tuple of pure strategies (one for each player) the Nash equilibrium requires that each player's strategy maximises her expected payoff against all other $n-1$ strategies.

The Nash Equilibrium plays a major role in modern economic theory, as *rational* players are assumed to maximise their payoff, and games will therefore typically evolve toward a Nash equilibrium. Other branches of economic game theory make different assumptions on what is optimised, and sometimes use different stability concepts. For instance, cooperative game-theory – where players are assumed to try to optimise the average payoff of all players in the game – uses the concept of "Pareto optimum", where no player can increase her payoff without decreasing the payoff of another player. In the theory of bounded rationality (Simon, 1955, 1969), the consequences of limitations in knowledge are investigated, where players are not necessarily maximising, but rather *satisficing* their payoffs.

In evolutionary biology (after some pioneering work by R.C. Lewontin and W.D. Hamilton, as is discussed in Maynard Smith, 1982) the use of game theory took off with the work of Maynard Smith & Price (1973) and Maynard Smith (1982). Maynard Smith & Price introduced the concept of **Evolutionarily Stable Strategy** (ESS) in an analysis of the evolutionary advantages of "limited war" strategies in animal conflicts, such as the prober strategy introduced above. An ESS is a strategy that cannot be *invaded* by any other strategy, because all other strategies get either a lower payoff when playing against the ESS, or if their payoff is equal, they get a lower payoff when playing against themselves. That is, if $F(i,j)$ gives the payoff for a player playing strategy i against an opponent playing strategy j, then i is an ESS if for every strategy j either $F(i,i) > F(i,j)$ or $F(i,i) = F(i,j) > F(j,j)$. Every ESS also defines a Nash Equilibrium, but the stability criterion is stricter, because it implies that every alternative strategy will be selected against if it occurs at small but non-zero frequency in the population.

In the example of figure 2.5, we can see that the dove-strategy is not an ESS, because the hawk-strategy has a higher payoff when playing against it. In a populations of doves, the hawk

[3]Grafen (2003) attributes the discovery of the Nash equilibrium to William Waldegrave, 1713, and refers to A. Hald (1990), "A History of Probability and Statistics and Their Applications before 1750", New York: Wiley Interscience.

strategy thus enjoys an initial selective advantage and will increase in frequency. The hawk-strategy is not an ESS either. A population consisting of just hawks can in turn be invaded by the dove-strategy, which has a higher payoff in a population of hawks, or by the prober-strategy, which has equal payoff against hawk but a higher payoff against itself. Only the prober strategy, in the present simple model, is an ESS: both doves and hawks fare worse than the prober in a population of probers[4].

If we exclude the prober-strategy from the strategy set, the resulting hawk-dove game has no ESS, i.e. a population of individuals all playing one pure strategy, can be invaded by the other strategy. In such games there might still be a stable distribution of phenotype frequencies in a population – called an **Evolutionarily Stable State**. In such a situation, there are distinct, genetically different players in the population ("polymorphism"), and this polymorphism is maintained by selection. Interestingly, such a stable distribution with p doves and $1 - p$ hawks is equivalent to a population where each individual plays the dove-strategy with *probability* p and the hawk-strategy with probability $1 - p$. If such *mixed strategies* are included in the strategy set (that is, allowed according to the hypothesised constraints), it is an ESS[5] and there is no polymorphism maintained.

The techniques and formalisms from evolutionary game theory immediately lead to some fundamental observations on the evolution of communication. Consider the evolution of an alarm call system similar to the calls that, for instance, ground squirrels (Sherman, 1977) or vervet-monkeys (Seyfarth *et al.*, 1980) use to inform conspecifics of the presence of predators. If we focus on just two signals, 1 and 2, and just two types of predators, aerial (E, e.g. eagles) and terrestrial predators (L, e.g. leopards), we can postulate the following strategy set:

Sender	A:	send 1 when observing E; send 2 when observing L.
strategies	B:	send 2 when observing E; send 1 when observing L.
	C:	never send anything.
Receiver	A':	act as if observing E when hearing 1; act as if observing L when hearing 2.
strategies	B':	act as if observing E when hearing 2; act as if observing L when hearing 1.
	C':	ignore all received calls.

[4] In the original paper (Maynard Smith & Price, 1973), this game was introduced with "dove" labeled "mouse" and "prober" labeled "prober-retaliator". Incidentally, an unfortunate choice of parameters resulted in there being in fact no ESS at all, even though a fourth strategy "retaliator" was erroneously identified as such.

[5] Grafen (1979) points out that mixed strategy ESS's and pure strategy evolutionary stable states are not equivalent in kin selection models.

In the case of alarm calls, the payoffs for sender and receiver are very different. The sender will suffer a cost, because by calling she alerts the predator of her presence and location. Evidence of the existence of a real cost in nature comes from the fact that alarm calls typically have very high pitch, which makes it more difficult for predators to locate the caller (Maynard Smith, 1982). The payoff matrix for the sender will therefore have all negative entries (parameter c) for strategies A and B, and (by definition) 0 for strategy C.

The receiver, on the other hand, will profit from a call *if and only if she correctly interprets it.* That benefit is quantified with parameter b. If the actual predator is a leopard, acting as if an eagle is observed can be a costly mistake: monkeys flee into the bushes to escape from an eagle attack, but that is in fact exactly where leopards hide (Seyfarth & Cheney, 1997). The cost of mis-interpretation is quantified as parameter m. If the receiver ignores all calls, her payoff is 0 (again, by definition). The payoff matrices in this simple example will thus look as in figure 2.6.

	receiver strategy		
sender strategy ↓	A'	B'	C'
A	$-c$	$-c$	$-c$
B	$-c$	$-c$	$-c$
C	0	0	0

(a) sender's payoff

	receiver strategy		
sender strategy ↓	A'	B'	C'
A	$+b$	$-m$	0
B	$-m$	$+b$	0
C	0	0	0

(b) receiver's payoff

Figure 2.6: Payoff matrices in a simple alarm call system

It is clear that neither A nor B can be the stable strategy for the speaker; if the cost of calling, c, is non-negligible, the strategy of not communicating at all, C, is always optimal. In explaining the evolution of communication, we thus face a **problem of cooperation**: if the benefits of communication are for the hearer, the sender has no incentive to give away her information, or even put herself at risk. Dawkins & Krebs (1978) pointed out this problem with what they call the "classical ethological" view on animal communication, which takes communication as existing for the benefit of the group. Dawkins and Krebs have therefore suggested that communication should be understood as a form of manipulation, with the benefits of successful manipulation with the sender.

Others (e.g. Maynard Smith, 1965; Sherman, 1977; Cavalli-Sforza & Feldman, 1983) have argued that "altruistic" communication can evolve through kin selection. However, the appropriateness of kin selection for human language – where communication is typically with non-kin – has been called into question (Dessalles, 1998). Dessalles has instead argued for a form of "reciprocal altruism", where there is a real benefit for the sender, because it is rewarded with status

in the population. Fitch (2004) reviews his and other arguments, but concludes that they are not convincing. He posits the "mother tongue" hypothesis – that human language developed primarily in a context of kin communication – as one of a number of factors that shaped human language in its evolution, and calls for further exploration of the role of kin selection in language evolution.

In many circumstances, for instance sexual signaling, the problem is not so much in the willingness to send signals, because the senders benefit, but in the **honesty** of the signals. A large amount of work on the evolution of animal and human communication has been concerned with this problem, leading to what is now called "honest signaling theory" (the handicap principle, Zahavi, 1975, 1977; Grafen, 1990). Hence, the problem of cooperation is pervasive in work on the evolution of communication, although its instantiations differ with different assumptions on the costs and benefits of communication, for both sender and receiver. Although the problem of cooperation is a consequence of careful considerations of payoff, strategy sets and invasibility, I will, because of its importance, add it as a separate point to the list of requirements of evolutionary explanations:

Criterion 8 (Problem of cooperation) *Evolutionary explanations of the evolution of language need to address the problem of cooperation, and demonstrate that senders will be willing to send honest signals, and that hearers will be willing to receive and believe the signal.*

Even if we find a scenario where successful communication is in the interest of both the speaker and the hearer, there is another problem that arises from the frequency-dependence of language evolution. We could call this the **problem of coordination**. If we ignore the non-cooperative strategies C and C', how does a population of players coordinate their behaviours such that they play either A and A', or B and B'? That is, how do they agree on a shared code? This problem seems particularly difficult when we consider a series of innovations, as in Jackendoff's (2002) scenario of the evolution of human language (chapter 3). Each of these innovations needs to confer a fitness advantage if it is to spread the population, but it is difficult to see how a genuine innovation can be advantageous to the individual if it is not shared by the rest of the population (Zuidema & Hogeweg, 2000; Zuidema & de Boer, 2003, see appendix C of this thesis).

Lewis (1969) showed that only "perfect" communication systems are "separating equilibria", which, if the role of "rationality" of the players is replaced by natural selection, corresponds to evolutionary stable states (Skyrms, 1996; Trapa & Nowak, 2000; van Rooij, 2004). Models in this tradition make the following assumptions:

- There is no cost to communication;

- The interests of sender and receiver are perfectly aligned;
- There is a discrete set of signals and a discrete set of meanings, and the number of signals equals the number of meanings;
- All meanings are equally frequent and valuable;
- Every "perfect" mapping from meanings to signals is equally good (which implies that meanings have no relation to each other, signals have no relation to each other, and meanings have no natural relation to signals);
- The meaning–signal associations are innate and inherited from parent to child.

It is easy to see why perfect communication systems are the only ESS's under these assumptions: if a communication system is sub-optimal, there must be synonymy: multiple signals are used for the same meaning. For the sender, however, it is always best to express a meaning m with the single signal s that has the highest chance of being understood, i.e. to avoid synonymy. The alternative signal(s) will thus not be used to express m anymore, and becomes available (through drift) for meanings that cannot be expressed yet. Hence, only "perfect" systems are stable against selection and drift.

It is clear, however, that all of these assumptions are violated in reality. Signals do have a cost, interests are not perfectly aligned, meanings and signals are not discrete, symbolic entities, but have similarity relations with themselves and each other, and, at least in human language, meaning–signal mappings are learnt and not innate. The problem of coordination thus remains a major open issue in the evolution of language, which we can add to the list of requirements:

Criterion 9 (Problem of coordination) *Explanations for the evolution of language need to deal with the problem of coordination, that is, show how, after each innovation, a shared code can be established and maintained.*

Much of the work on the evolution of language can be seen as dealing with this problem. A number of models, for instance, relax the innateness assumption above, and study, in computer simulations, the evolutionary success of a number of different strategies in word learning (Hurford, 1989; Oliphant, 1999; Smith, 2004). The payoff function in Hurford's model is the expected success in communication between a sender and a receiver (i.e. the game is cooperative; both sender and receiver benefit from success). Sender behaviour is characterised by a probabilistic mapping from a set of M meanings to a set of F signals; receiver behaviour by a probabilistic mapping from the signals to the meanings.

Hurford was interested in how these functions were learnt, and in the evolution of different learning strategies. The strategy set Hurford considered consisted of three strategies, termed

imitator (that imitates the observed average sending and receiving behaviour in the population), calculator (that estimates the best send and receive functions based on observations of the population's receive and send behaviour respectively) and Saussurean learner (that chooses the same receive function as the calculator, but derives the send function from that receive function rather than from the receiving behaviour in the population). Hurford showed that Saussurean learners outcompete the other two learning strategies. These results were extended by Oliphant & Batali (1996), Oliphant (1999) and Smith (2004), among others. From these studies it emerged that learning strategies can evolve that give rise to "perfect" communication systems in a population.

The model I will study in chapter 5 and related work (e.g. Nowak & Krakauer, 1999), does not model such explicit learning rules, but does relax some of the other assumptions mentioned. I report results where the number of signals is larger than the number of meanings, where there is noise on signals, where some meanings are more valuable than others and where there are similarity relations between meanings and between signals. More work is needed to study whether these results hold when learning is modelled explicitly. An encouraging result in this respect is due to Calvin Harley (1981). He studies the evolution of learning rules and showed that evolution will favour rules that *learn* the evolutionary stable strategy. Hence, results on Evolutionary Stable Strategies in innate communication systems, in principle carry over to situations where the same strategies are acquired in a learning process (Maynard Smith, 1982, chapter 4).

2.8 Levels of Selection

I have discussed some basic concepts from population genetics, which describes the change in frequencies of *genes*, and from evolutionary game theory, which describes the invasion and replacement of phenotypic *strategies* of individuals. The two approaches are obviously related, because the fitnesses of genes are dependent on the phenotypes they code for, and a strategy will only replace another strategy if all the genes necessary for that strategy are selected for and get established in a population. But the description of the evolutionary process in population genetics and evolutionary game theory are set at entirely different levels.

In Dawkins' (Dawkins, 1976) terminology, genes are *replicators*: they are the bits of information that get copied and transmitted – more or less intact – to the next generation. Individuals are *vehicles* (Dawkins, 1976) or *reproducers* (Szathmáry, 1999). In sexual species, such as humans, a child is radically different from any one parent, because she inherits only 50% of the genes. Individuals, therefore, are not replicators, even though they are the obvious level of description when we talk about fitnesses and strategies.

If *replicators* and *reproducers* were the same objects, evolutionary dynamics would be relatively easy to describe. But in general, especially in sexual species, they are not. Genes are "packaged" – contained within the structured genome of an individual that lives within a structured population. That packaging makes the fate of a specific gene depend on the other genes it is associated with (genes that occur together more often or less often than would be expected on the basis of their frequencies alone, are said to be in *linkage disequilibrium*). If a gene a happens to be associated with a gene b that is under strong positive selection, gene a will increase in frequency even though it does not itself contribute to the fitness of its carrier ("genetic hitch-hiking", Hill & Robertson, 1966; Maynard Smith & Haigh, 1974). To predict the fate of a specific gene, we therefore need to know the statistical associations with other genes.

To make things even more complicated, not just the gene frequencies change; also the associations themselves change in evolution. The *physical linkage* between genes on a chromosome tends to keep these genes together, but *recombination* breaks up these associations; *sexual selection* generates associations between for instance, the preferences of the females and the selected traits of the males; finally, *epistasis* also generates linkage equilibrium, because if genes are much better in combination than they are apart, natural selection itself will make the combination more frequent than expected by chance. Barton & Turelli (1991) and Kirkpatrick, Johnson & Barton (2002) have developed a mathematical framework to describe the dynamics of such *multi-locus evolution*; however, they take fitnesses as given and do not yet provide a bridge to the fitness concept in phenotypic models.

Hence, the relation between gene frequency change and adaptation at the level of the individual (such as language) is not at all trivial. The problem with the gene as the level of description is that we don't know the relevant fitness coefficients, because our knowledge of life, death and reproduction is almost entirely specified at the level of the individual. But the problem with the individual as level of description, is that we are not necessarily justified in assuming that natural selection corresponds to optimisation. Do the results from game-theoretic analyses translate to fitness coefficients of the genes that underlie the strategies? How do we relate the fitness coefficients, and the fundamental results about evolution as optimisation by Fisher and Wright, to adaptation on the level of individuals? Grafen (2003), in a discussion of Fisher's "fundamental theorem of natural selection" (Fisher, 1930) observes that (too) few researchers in evolution worry about these issues:

> *"the theorem was fundamental in 1930 because it isolated the adaptive engine in evolution and made an extraordinary link between gene frequencies and adaptive*

change. It really did show how Darwinian natural selection worked simply and consistently and persistently amid the maelstrom of complexities of population genetics. The theorem is just as important today for that reason. This is not popularly realised by biologists because most take for granted an informal sense that natural selection leads to organisms maximizing their fitness, but they do not ask how that sense can be justified." (Grafen, 2003, p.327)

Grafen lists three assumptions that are made in the original version of Fisher's theorem, and apply equally to Wright's equations discussed in section 2.4:

- It assumes the fitnesses of genes are frequency independent. That is, the fitness of a given genotype is not dependent on which other genotypes are present and at which frequencies in the population. Consequences of frequency dependence are studied in evolutionary game-theory (Maynard Smith & Price, 1973; Maynard Smith, 1982).
- It assumes that all individuals interact with all other individuals with equal probability. That is, it assumes the fitness of a given genotype is not dependent on the genotypes which are potentially correlated with it. Consequences of such correlations are studied in social evolution theory (Hamilton, 1964a,b; Frank, 1998).
- It assumes fitnesses are fixed; Grafen himself has worked on the consequences of natural selection under uncertainty.

For the purposes of this chapter, it would take too far to investigate the contributions of Grafen and others to relate population genetics and evolutionary game theory. However, a few important implications for language evolution research from the discussion so-far are worth making explicit. First, a "strategy" in a game-theoretic analysis will typically be coded for by many genes (*pleiotropy*). So if alleles $a_1, a_2 \ldots a_n$ at loci 1 to n are needed for an evolutionarily stable strategy A, we need each of these alleles to represent a step in the right direction. In technical terms, we need *additive genetic variance*; Maynard Smith (1982) argues that additive genetic variance is common in nature, and that this is therefore a reasonable assumption to make in game-theoretic analyses. We need to be aware, however, that we ignore all the phenomena of multi-locus evolution in game-theoretic analyses of language.

Criterion 10 (Levels of selection) *Explanations for the evolution of language need to relate selection at the level of individuals or groups to changes in gene frequencies. That is, they need to specify and relate the assumed levels of description for selection and heritability.*

Second, an important (methodological) observation is that there is no single best level of description; researchers make a heuristic choice about the level at which they will describe the evolutionary dynamics. Every model will only be an approximation, and it depends on the phenomenon of interest at which level the evolutionary process is most adequately described. Below, I will briefly discuss kin selection, and show, using the Price equation, why for the phenomena of social evolution the population structure is a crucial level of description that is left out in standard game-theoretic models.

2.9 Social Evolution & Kin Selection

The techniques from social evolution theory will not play much of a role in this thesis, and I will therefore keep the discussion brief. One fundamental equation, the **Price equation** (Price, 1970), is useful, however, to highlight a silent assumption in game theory models, and to illustrate the issue of multiple levels of selection. The Price equation is easily derived; I will follow here Frank (1998) and Andy Gardner (p.c.). Like Wright's equation (2.9), it can be interpreted as describing the change in the frequency of a gene, but more generally it describes the change in the value of any trait z.

Price introduces his equation as follows:

> *"Gene frequency change is the basic event in biological evolution. The following equation [...], which gives frequency change under selection from one generation to the next for a single gene or for any linear function of any number of genes at any number of loci, holds for any sort of dominance or epistasis, for sexual or asexual reproduction, for random or nonrandom mating, for diploid, haploid or polyploid species, and even for imaginary species with more than two sexes" (Price, 1970, p.520)*

We are interested in the change in frequency of a specific trait z in the population between the present ($\overline{z}$) and the next generation ($\overline{z}'$). If we divide up the population in M units $q_1 \dots q_M$ (these units are, for instance, individuals or groups, depending on the level of selection the equation is meant to describe), and we know their fitnesses $w_1 \dots w_M$ and trait values $z_1 \dots z_M$, then the change of the trait's frequency in the whole population is given by:

$$\begin{aligned} \Delta\overline{z} &= \overline{z}' - \overline{z} \\ &= \sum_i q_i' z_i' - \overline{z} \\ &= \sum_i q_i \frac{w_i}{\overline{w}} (z_i + \Delta z_i) - \overline{z} \end{aligned} \tag{2.12}$$

Multiplying both sides of this equation with $\overline{w}$, and rearranging gives:

$$\begin{aligned} \overline{w}\Delta\overline{z} &= \sum_i q_i w_i z_i + \sum_i q_i w_i \Delta z_i - \overline{w}\,\overline{z} \\ &= \underbrace{\sum_i q_i w_i z_i - \overline{w}\,\overline{z}}_{\mathrm{Cov}[w,z]} + \underbrace{\sum_i q_i w_i \Delta z_i}_{E[w\Delta z]} \end{aligned} \tag{2.13}$$

As indicated, the terms in equation (2.13) correspond, by definition, to the *covariance* between fitness and trait value, and *expected value*[6]. Hence, the process of evolution can be elegantly summarised in the Price equation, as follows:

$$\overline{w}\Delta\overline{z} = \underbrace{\mathrm{Cov}[w,z]}_{\text{selection}} + \underbrace{E[w\Delta z]}_{\text{transmission}} \tag{2.14}$$

The Price equation partitions the process of evolution into a term that describes the effects of selection (traits that are associated strongly with fitness will be selected for most effectively), and a term that describes the effects of (biased) transmission (the index i is the index of the parent; hence Δz_i describes the change in the trait value – from a particular parent to all its offspring – regardless of selection).

Observe that the transmission term in the Price equation looks very similar to the left-hand side of that equation. This fact allows us to relate different levels of selection. As an illustration, I will here derive Hamilton's (Hamilton, 1964a,b) famous result on kin selection, which says that an altruistic trait can evolve if the benefit b times the relatedness r is larger than the cost c:

$$br > c. \tag{2.15}$$

The derivation using the Price equation highlights the correct interpretation of *relatedness* and suggests applications for language evolution. The derivation concerns the evolution of an altruistic trait, such as the alarm calls discussed in the previous section. For simplicity, assume an individual either does or does not have this trait. We indicate this with the variable z, that is, $z = 1$ or $z = 0$. We can ask: under which circumstances will this trait evolve?

[6]The covariance between two variables x and y is defined as $\mathrm{Cov}(x_i, y_i) = \frac{1}{N}\sum_{i=1}^{N}((x_i - \overline{x})(y_i - \overline{y})) = \overline{xy} - \overline{x}\,\overline{y}$, i.e. the product of the means minus the mean of the products. Expected value of a variable x is defined as $E(x) = \sum_{i=1}^{N} P(x = x_i)x_i$, i.e. the sum of all possible values weighted by the probability of each value. Covariance is the most obvious way of measuring a departure from statistical independence. If x and y vary independently from each other, then $E(xy) = E(x)E(y)$, and the covariance is 0.

Consider a population, subdivided (at random) in N groups $G_1 \ldots G_N$, each of size M individuals. In each group G_i, individuals benefit from the amount of altruism in that group, labelled as z_i; the total benefit is bz_i. The jth individual in that group, however, also suffers a cost from being altruistic, indicated with c; the cost is thus cz_{ij}. The fitness of the jth individual in the ith group is now given by:

$$w_{ij} = \alpha + bz_i - cz_{ij}, \tag{2.16}$$

where α is a baseline fitness (not dependent on the presence or absence of the altruistic trait). The fitness of the ith group is given by:

$$w_i = \alpha + (b - c)z_i. \tag{2.17}$$

Hence, an individual's fitness (her relative contribution to the total offspring of the group) depends on the amount of altruism received and the amount of altruism given. Obviously, if the cost c of being altruistic is larger than 0, it is always best for an individual to be selfish. The group's fitness[7] (the relative contribution of this group's offspring in the total offspring of the whole population) depends on the total amount of altruism given. If the cost c of altruism is lower than the benefit b, it is always best *for the group* if all individuals are altruistic.

The evolutionary process within each group i can be described with a Price equation, as in equation (2.14). If we assume there is no transmission bias, the equation simplifies to:

$$\overline{w_{ij}}\Delta\overline{z_{ij}} = w_i\Delta z_i = \mathrm{Cov}_j[w_{ij}, z_{ij}]. \tag{2.18}$$

The evolutionary process at the level of the whole population is also described with a Price equation, where the transmission term concerns the within-group dynamics of equation (2.18):

$$\begin{aligned} \overline{w_i}\Delta\overline{z_i} &= \mathrm{Cov}_i[w_i, z_i] + E_i[w_i\Delta z_i] \\ &= \mathrm{Cov}_i[w_i, z_i] + E_i[\mathrm{Cov}_j[w_{ij}, z_{ij}]]. \end{aligned} \tag{2.19}$$

The covariance in above equation can be replaced by a regression and variance term, because (by definition) $\mathrm{Cov}(x, y) = \beta(x, y)\mathrm{Var}(y)$. This gives the following equation:

$$\overline{w_i}\Delta\overline{z_i} = \beta(w_i, z_i)\mathrm{Var}_i[z_i] + E_i[\beta(w_{ij}, z_{ij})\mathrm{Var}_j[z_{ij}]]. \tag{2.20}$$

[7] Note that, although parent groups are of fixed size M, some groups produce more offspring than others.

These regression terms β can be read off directly from equations (2.16) and (2.17), because they correspond to the slope of the fitness functions, i.e. $\beta(w_i, z_i) = b - c$ and $\beta(w_{ij}, z_{ij}) = -c$. Substituting these values into equation (2.20) and rearranging gives:

$$\begin{aligned}
\overline{w_i}\Delta\overline{z_i} &= (b-c)\mathrm{Var}_i[z_i] + E_i[-c\mathrm{Var}_j[z_{ij}]] \\
&= (b-c)\mathrm{Var}_i[z_i] - cE_i[\mathrm{Var}_j[z_{ij}]] \\
&= b\mathrm{Var}_i[z_i] - c(\mathrm{Var}_i[z_i] + E_i[\mathrm{Var}_j[z_{ij}]]) \\
&= b\mathrm{Var}_i[z_i] - c\mathrm{Var}_{\text{total}} \\
&= \left(b\frac{\mathrm{Var}_i[z_i]}{\mathrm{Var}_{\text{total}}} - c\right)\mathrm{Var}_{\text{total}},
\end{aligned} \tag{2.21}$$

where $\mathrm{Var}_{\text{total}}$ is the total variance. This establishes a derivation of Hamilton's rule from the Price equation, because the average relatedness between two individuals in a population, equals the between group variance as a proportion of the total variance. That is, $r = \frac{\mathrm{Var}_i[z_i]}{\mathrm{Var}_{\text{total}}}$. If the benefits of trait z, weighted with the relatedness within a group, are larger than the costs, i.e. $rb > c$, then $\Delta\overline{z}$ will be positive, i.e. evolution will favour the trait even if it harms the individual.

It is important to note that Hamilton's rule is widely misinterpreted. As this derivation shows, the relatedness term r is *not* the fraction of genes two individuals share (*identity by descent*), as is commonly assumed (e.g. Okasha, 2003). Rather, it is a statistical association between the trait of interest in one individual and the trait in the individual she interacts with. Therefore, the relatedness between two individuals can even be negative. This simply means that the individuals are less related to each other than to a random third individual in the population (Hamilton, 1970). If the association is high enough, altruistic traits can be favoured by natural selection[8]. That is, if (for whatever reason) altruists are surrounded by other altruists, they benefit more from the altruism received than from the altruism offered (and conversely, if it is low enough, natural selection can favour *spite* – behaviours that harm both the actor and the recipient; Hamilton, 1970; Gardner & West, 2004).

[8]Darwin already understood the essence of kin selection when he wrote: ""[...] selection may be applied to the family, as well as to the individual, and may thus gain the desired end. Thus, a well-flavoured vegetable is cooked, and the individual is destroyed; but the horticulturist sows seeds of the same stock, and confidently expects to get nearly the same variety; breeders of cattle wish the flesh and fat to be well marbled together; the animal has been slaughtered, but the breeder goes with confidence to the same family. [...] Thus I believe it has been with social insects: a slight modification of structure, or instinct, correlated with the sterile condition of certain members of the community, has been advantageous to the community: consequently the fertile males and females of the community flourished, and transmitted to their fertile offspring a tendency to produce sterile members having the same modification." (Darwin, 1859, p.258-259)

Interactions within kin-groups (and kin recognition) are an important mechanism for this association to arise (hence the Maynard Smith's term "kin selection"), but not the only one. Subdivision of a population in groups is another mechanism (such "group selection" is thus a form of kin selection). Hamilton himself suggested a third mechanism, that of "green beards". If the same gene complex that codes for an altruistic trait, also codes for an external marker (i.e. a green beard), altruists can choose to interact preferentially with each other. This is of interest for language evolution, because language itself could be such a green beard, if individuals with a linguistic innovation can recognise each other based on features in their language. Finally, reciprocal altruism (Trivers, 1971), where players remember the interaction history with other players and play altruistically only against players that have been altruistic in the past, can be understood in the same framework.

Kin selection seems the most promising solution for the problem of cooperation that I introduced in section 2.7. It would certainly be worthwhile to study formal models of kin selection, that take into account the details of human communication. In this thesis, however, I will no further address kin selection or the problem of cooperation. Instead, I will assume the willingness to cooperate exists in modeled populations, and focus on the problem of coordination.

2.10 Cultural Evolution

Dawkins (1976) emphasised that the principle of natural selection is not restricted to genes or individuals (as Fisher, Wright, Haldane, Price, Hamilton and others were well aware). In every situation where one can identify replicators, heritable variation and natural selection, a process of adaptation can take place. For instance, cultural inventions (or "memes", Dawkins, 1976) – religion, technology, fashion or indeed words and grammatical rules – undergo evolution if there are mechanisms for cultural transmission and cultural selection.

Since Dawkin's book, many wildly speculative theories have been launched under the heading "memetics", which have given this new field a bad reputation. Nevertheless, the basic idea is sound and open to serious investigation (Mesoudi *et al.*, 2004). For a start, all mathematical models and requirements discussed in this chapter apply, mutatis mutandis, to cultural evolution as well. The idea of viewing historical language change as a form of evolution is particularly attractive because, on the one hand, it makes the extensive mathematical toolkit of evolutionary biology available to linguistics, and on the other hand, is presents evolutionists with an enormous body of new data. In chapter 7 I will briefly come back to some possible implications for linguistics and biology.

We need formal models of the cultural evolution of language, in which we can deal with all the constraints on evolutionary models that I listed in this chapter. Although many authors have noted the parallels between biological evolution and language change, including Darwin (1871, p.91), only recently have people starting to study the cultural evolution of language in such a formal framework. Some relevant mathematical models are those of Cavalli-Sforza & Feldman (1981), Niyogi (2002) and Yang (2000). These authors look at the competition between two or more languages, with no qualitative differences between languages. Simulation models such as those of Kirby (1998) and Batali (2002) look at more open-ended systems, with more explicit formalisms for grammar and learning.

One problem is that is not so easy to decide on the appropriate units of selection. For instance, Kirby (2000) described the dynamics in his simulation model (which will be discussed in detail in chapters 5 and 6) with context-free grammar rules as replicators under selection for more reliable replication. In later papers, however, he argued that the analogy between biological and cultural evolution in this case breaks down (Kirby, 2002b). This is because the grammatical rules are *induced* from observable language, whereas in biological evolution genes are *inherited*, with no feedback from phenotype to genotype (other than through the effects of selection). This is known as the "central dogma of molecular biology". This observation is correct, of course, but it does not mean we cannot describe the dynamics in models such as Kirby's using the tools from evolutionary biology. The effects of induction in language change are a form of "directed mutation", and can be included, for instance, in the Price Equation in the transmission term. More work is needed to work this out with concrete examples; chapter 6 will attempt to make a small contribution to this end.

2.11 Conclusions

In this chapter I have discussed a variety of models from population genetics, evolutionary game-theory and social evolution theory. I have used these models to make a list of requirements for evolutionary scenarios of the biological and cultural evolution of language. These requirements correspond to the following questions we should ask when confronted with a scenario for the biological or cultural evolution of language:

- What are the units of inheritance the scenario assumes? Genes? Memes?
- What is the scope of variation in these genes or memes? That is, what is the assumed set of possible traits/strategies available for evolution?

- What are the selection pressures? That is, what is the assumed payoff for each of these possible traits in each possible context?
- For every innovation in the scenario, will it indeed be favoured by selection when extremely rare? If not, is there a non-negligible chance it could get established by stochastic effects, or get frequent enough to be favoured by selection?
- Does the assumed series of changes in the scenario indeed constitute a path of ever-increasing fitness? That is, is there a path of fit intermediates from start to finish?
- How much time will each of the innovations take to get established?
- Is there for every transition sufficient variation, but not too much?
- How does the scenario explain that speakers maintain the willingness to speak honestly, and that hearers continue to listen and believe the information received? That is, how does it solve the problem of cooperation?
- How does the scenario explain that speakers and hearers, after every innovation, agree on which signals refer to which meanings? That is, how does it solve the problem of coordination?
- How does the scenario relate dynamics at different levels of description – genes, strategies, individuals, groups, languages?

With these questions in mind, the next chapter will discuss a possible scenario for the evolution of language, proposed by linguist Ray Jackendoff (2002). I will identify the transitions in the scenario that can be addressed in this framework, and study these transitions in the rest of the thesis.

CHAPTER 3

The major stages in the evolution of language

What are the "design principles" of human language that need to be explained? In this chapter I will discuss Jackendoff's scenario for the evolution of language, and argue that the transitions between stages are the crucial challenges for evolutionary biology. I will review a number of formalisms for meaning, sound and the mapping between them, and describe and evaluate the differences between each of Jackendoff's stages in terms of these formalisms. I conclude from this discussion that the transitions to combinatorial phonology, compositional semantics and hierarchical phrase-structure can be formally characterised. Modelling these transitions is a major challenge for language evolution modelling.

3.1 Introduction

Human languages are unique communication systems in nature because of their enormous expressiveness and flexibility. They accomplish this by using combinatorial principles in phonology, morphology, syntax and semantics (Chomsky, 1955; Studdert-Kennedy, 1998; Jackendoff, 2002), which impose important requirements on the cognitive abilities of language users. Explaining the origins of the structure of language and the human capacity for learning and using it, are challenging and controversial problems for linguistics, cognitive science and evolutionary biology. Major disagreements concentrate on whether or not this capacity has been subject to natural selection or not, whether it evolved in a single, in few or in many steps, and whether articulation, perception or cognitive processing formed the crucial bottleneck (see Christiansen & Kirby, 2003a, for a representative overview of current positions).

Jackendoff (2002) has laid out a scenario for the various stages in the evolution of human language from primate-like communication. Unlike many other theories, Jackendoff's scenario assumes many such intermediate stages, as summarised in figure 3.1. Jackendoff's proposal is useful for structuring the discussion in this thesis for a number of reasons:

- Jackendoff's scenario is a gradualist account, with many intermediate steps. Scenarios proposed by other scholars can be seen as variants of Jackendoff's, where two or more of the stages Jackendoff proposes are collapsed into one. Jackendoff's scenario can thus be seen as a generalisation of many other scenarios.
- It grounds the scenario for the evolution of language in a testable theory of how modern humans acquire, represent and process language. Although Jackendoff's account is not very formal, his partitioning of the problem is useful for identifying the relevant formalisms from modern linguistics and applying them to issues in language evolution.

Jackendoff hypothesises that modern languages still contain elements that correspond to the type of elements that characterised earlier stages in the evolution of human language. For instance, he views the compound noun construction in English as a *fossil* of an earlier stage where strings are concatenated to express a compound meaning, but without recursive phrase-structure. Thus, the meaning of compounds like "dog house" and "house dog" is deducible (but not completely specified) from the meaning of the component words and the order in which they are put.

Although an incremental, step-by-step scenario is a crucial component of an evolutionary theory, Jackendoff's scenario does not address the other crucial component: the transitions from each stage to the next. Jackendoff admits: *"I will not inquire as to the details of how increased*

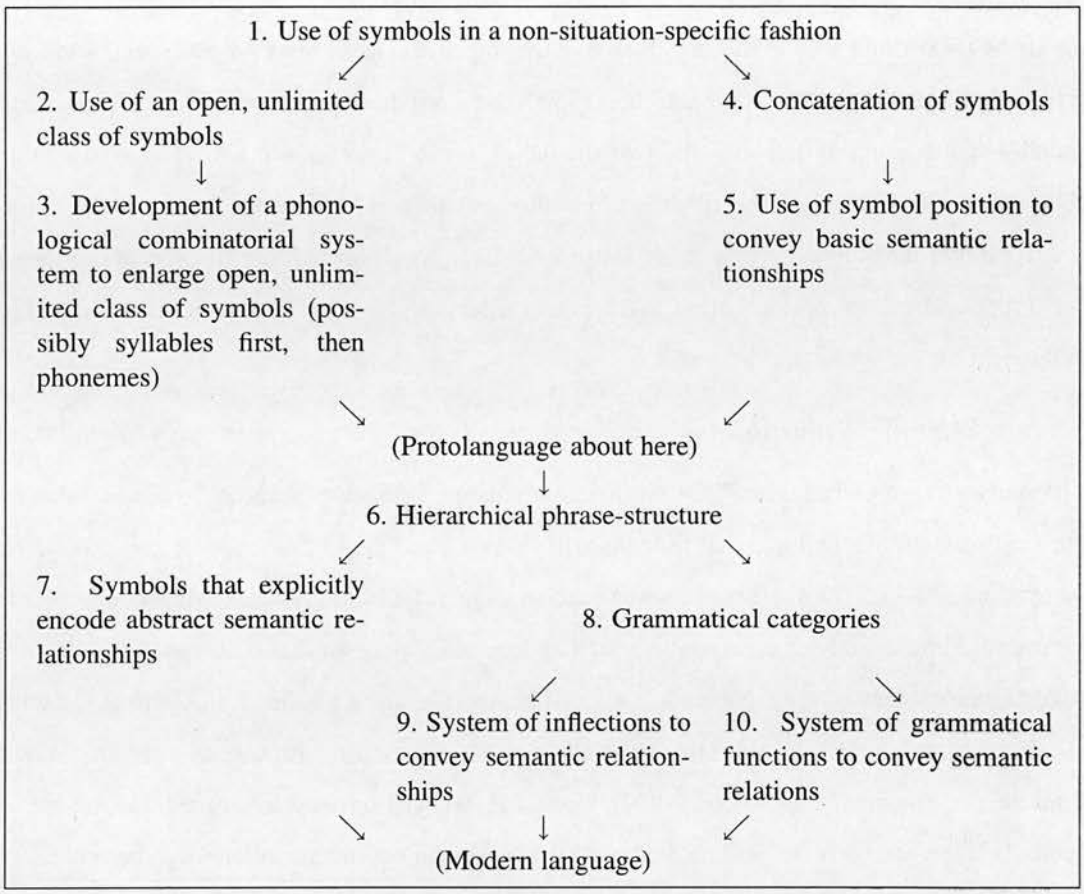


Figure 3.1: Incremental evolutionary steps in Jackendoff's scenario for the evolution of human language (from Jackendoff 2002). Independ steps appear side by side; dependencies among steps are indicated vertically.

expressive power came to spread through a population [...]. nor how the genome and the morphogenesis of the brain accomplished these changes. Accepted practice in evolutionary psychology [...] generally finds it convenient to ignore these problems; I see no need at the moment to hold myself to a higher standard than the rest of the field." (Jackendoff 2002, p. 237)

Presumably, all transitions have greatly increased the number of distinct "signs" (signal–meaning pairs) that can be expressed, transmitted, memorised and learnt. However, that observation in itself is not sufficient. As I have argued in the previous chapter, good evolutionary explanations must specify the selection pressures that lead from one stage to the next (the payoff function), and the variation that natural selection can act upon (the strategy set). Understanding how innovations can spread in a population is the essence of any evolutionary explanation, and, crucially, a better end-result is neither a sufficient nor a necessary condition for the spread of innovations.

In the following I will briefly sketch each of the stages that Jackendoff proposes, and relate his proposal to those of some other researchers. I will then introduce some formalisms for meaning, sound and the mapping between the two (including formalisms for syntax). The goal of this chapter is to find out how to describe the similarities and differences between Jackendoff's stages, and to identify the transitions that can be formally characterised. In the other chapters of this thesis I will then address possible solutions to the more difficult problems of how to get from one stage to the next.

3.2 Jackendoff's Scenario

The starting point of Jackendoff's scenario is pre-existing primate conceptual structure – that is, the kind of cognitive abilities that modern primatology has found in other great apes. The first innovation is the **use of symbols** in a non-situation specific fashion. Jackendoff recognises that for instance chimpanzees have a sophisticated conceptual apparatus, that is adequate to deal with navigation, problem-solving and complex social interaction. But he believes that primates are incapable of symbolic vocalisation, that is, of referring to objects, properties, or events independent from the present situation. Deacon (1997), Donald (1991) and others have argued that the use of symbols is the crucial biological innovation that has made modern human language possible.

A second innovation is the ability to use and acquire an **open, unlimited class of symbols**. Whereas primate call systems contain a few dozen different calls at most (as far as we know) and language-trained apes can be taught at most several hundred symbols, humans know tens of thousands of different words. An open class of symbols must be learnt rather than be innate; Oliphant (1999) and others have argued that a learnt vocabulary was a crucial step in the evolution of language.

To keep such large numbers of symbols distinct in perception, memory and articulation, a third innovation has been crucial: a **generative, combinatoral phonological system**. All human languages are combinatorial, in that the many basic meaningful units (words, morphemes) are built up from a relatively small repertoire of basic speech sounds (phonemes, syllables). Jackendoff endorses the view that the syllable is the basic unit of combination. The evolution of combinatorial phonology is seen by a number of researchers as the crucial innovation in the evolution of language (Lindblom, MacNeilage & Studdert-Kennedy, 1984; Studdert-Kennedy, 1998).

Jackendoff's fourth innovation is the **concatenation of symbols** to build larger utterances. He imagines concatenations of symbols analogous to *"Fred apple"*, which might refer to any of a number of connections between Fred and apples. Although simple concatenation does not

fully specify the intended meaning, it is nevertheless, Jackendoff argues, more useful than single symbols in isolation.

The fifth innovation, however, **using linear position to signal semantic relations**, does introduce a systematic compositionality. In this stage of the scenario, simple principles such as "agent first", "focus last" and "grouping" could structure utterances analogous to *"dog brown eat mouse"*, such that it is clear that the brownness applies to the dog, the dog eats and the mouse is being eaten. In the terminology of Hurford (2000), the route from holistic to compositional language in this scenario is "synthetic", because compounds are *synthesised* from pre-existing meaningful signals (rather than that pre-existing holistic signals are *re-analysed* as being built-up from component parts).

Jackendoff sees the fourth and fifth innovations as independent from the second and third, and he does not decide which should come first (see figure 3.1). Together, they constitute something similar to the (single) intermediate stage of "protolanguage" in the scenario of (Bickerton, 1990) and others, and to pidgin (the limited language spoken by adults with different native languages, Bickerton, 1990) and to "Basic Variety" (the limited language acquired by adult second language learners, Klein & Perdue, 1997).

The sixth innovation is the invention of **hierarchical phrase-structure**. Phrase-structure has been recognised since Chomsky (1955, 1957) as a crucial design feature of human language. Jackendoff argues that phrase-structure allows the principles of word order, as emerged in stage 5, to be elaborated into principles of phrase order. Hence, from stage 6 a systematic relation has existed between sentences like *"dog chase mouse"*, where *"dog"* and *"mouse"* are single word noun phrases, and the similarly structured but more elaborate *"big dog with floppy ears and long scraggly tail chase little frightened mouse"*.

The seventh innovation is a **vocabulary for relational concepts**, introducing words analogous to *up*, *on*, *behind*, *before*, *after*, *often*, *because*, *however* and so forth. These words all describe relations between different phrases in a sentence, and thus require phrase structure, Jackendoff argues, but not yet syntactic categories. Jackendoff imagines that the phrase order and use of relational words are still completely guided by semantically defined notions. That is, there are no subjects, objects, nouns, verbs or mechanisms for case, agreement or long-distance dependencies. There are just semantic categories such as agent, patient, objects and actions.

Grammatical categories are the eighth innovation, creating syntactic notions such as "subject" that are correlated but not equal to the semantic notion of agent (as for instance in the passive construction), or even a syntactic notion like "sentence" which makes that *"a storm last*

night" cannot stand on its own, whereas *"There was a storm last night"*, with dummy subject *there*, can. The final two innovations, **inflectional morphology** and **grammatical functions** (in no particular order) complete the extensive toolkit that modern languages make use of. This list of gradual innovations is consistent with the gradualist approach championed by Pinker & Bloom (1990) and others.

In summary, Jackendoff breaks down linguistic competence into a number of different skills, and proposes a gradual scenario in which new skills are added to the existing system, each step increasing the expressivity of the language used. The first innovation, of symbol use, is about the sort of *meanings* early hominids had available for communication. The third, about combinatorial phonology, is about the kind of *sounds* they could produce and perceive. All the other innovations, from an open, learnt vocabulary and the concatenation of symbols to inflectional morphology, are about the way meanings are *mapped* onto sound and vice versa. In the next sections I will discuss some observations about and formalisms for the meaning, sound and meaning–sound mappings in animal and human communication. Where possible I will evaluate – in terms of those formalisms – whether Jackendoff's stages indeed capture the relevant innovations in the evolution of language. But as it turns out, it is often not straightforward to adapt the formalisms developed for describing modern, human languages to a different use, that is, to describing the differences between modern human communication and that of other species.

3.3 Modelling Meaning

Animals and humans categorise their environment, and use calls, words or grammatical sentences to express aspects of that environment. Typically, the same utterances are used on many different occasions to express common features. It is therefore reasonable to postulate an "internal state", a representation in the brain, that mediates between perceptual and motor abilities, memory, and linguistic forms. I will call these representations "meanings". Modelling the meaning of natural language utterances is difficult, because we have only very indirect access to the representations in the brain and, crucially, much of that indirect access is modulated through language (Hurford, 2003). The common framework for modelling meaning (that is, for formal semantics), is that of symbolic logic. Many different logics exists, with different levels of expressive power as well as different computational properties.

According to Jackendoff and others, the kind of meanings available for communication to modern humans are qualitatively different from those available to other primates, including our prelinguistic ancestors. Jackendoff believes that the "use of symbols" was the first major innovation; other researchers have argued that a "theory of mind" was a crucial innovation (Dunbar,

1998). It would be very useful if we could characterise such *conceptual* differences in formal terms, using the apparatus of formal semantics (that was developed with other purposes in mind). In the following I will briefly sketch a well-known hierarchy of logics (Chierchia & McConnell-Ginet, 1990; Gamut, 1991) – propositional, predicate and modal logic – to see if it can be used for such purposes.

Any discussion of logic, must start with **propositional logic** (or "Boolean algebra"), which provides a number of operators (such as negation '$\neg$', logical AND '$\wedge$', logical OR '$\vee$', implication '$\mapsto$' and bi-implication '$\leftrightarrow$') and inference rules that one can use to derive new true statements from other true statements (statements that can be true or false are called "propositions"). For instance, we can define the symbol p as denoting the statement *"Socrates is a man"*, and q as denoting *"Socrates is mortal"*. In propositional logic, the inference that q follows from p can then be described with the rule $p \mapsto q$. Using this rule, someone who is unsure whether Socrates is a real man or an imaginary figure (that is, the truth value of p), needs only confirmation that he is a man (p is true) to also infer that he is mortal (q is true).

Crucially, propositional logic does not have access to information "inside" a proposition. Thus, when we add r: *"Aristotle is a man"*, we have no way of generalising from Socrates' fate to Aristotle's and derive that *"Aristotle is mortal"*. **Predicate logic** does capture this generalisation, by introducing the notions of "predicates", "arguments" and "quantifiers" (such as the universal quantifier '$\forall$': *"for all"*, and the existential quantifier '$\exists$': *"there is"*). Hence, we can introduce the terms s for Socrates, and a for Aristotle, and the predicates H for being human, and M for being mortal. The fates of both wise men can now be inferred from two facts $H(s)$ and $H(a)$, and just one general inference rule $\forall x\,(H\,(x) \mapsto M\,(x))$, which is interpreted as *"All men are mortal."* That is, predicate logic can describe (and productively use) similarities between propositions that propositional logic has no access to.

Predicate logic is a powerful formal system that is used to model a good part of the semantics of natural languages as well as many other reasoning tasks in artificial intelligence. However, natural languages do contain many words and constructions with meanings that are difficult to model in predicate logic, such as modal constructions (*"can"*, *"may"*, *"must"*, *"perhaps"*, *"certain"*), tense (*"I will survive"*) and intentionality (*"I think that I am"*, where the content of the thought need not be true for the act of thinking it to be true).

It might appear that by introducing the right variables and predicates, predicate logic would be able to deal with such expressions. For instance, Kirby (2002a) uses (higher order) predicate logic notation for expressions such as *believes(pete,knows(gavin, hates(pete,heather)))* for *"Pete*

believes that Gavin knows that he hates Heather". Similarly, in the "flat" notation used by Batali (2002) and De Beule, Van Looveren & Zuidema (2002, see appendix C of this thesis) and advocated by Hurford (2003), we could introduce variables for events or situations, and predicates that define properties of these events such as *necessary*, *likely* or *possible*. *"John will always love Mary"* is then represented as something like:

$$\exists e\,(\mathit{loves}(e, x, y), \mathit{john}(x), \mathit{mary}(y), \mathit{always}(e))\,.$$

Such abuse of predicate logic notation might be useful for evolutionary simulations where the semantics is not really relevant. However, for characterising representational abilities it is not a real solution because it does not suggest a systematic way to evaluate the truth of expressions. David Lewis (1972, quoted in Abbott, 1999) makes this point as follows in a critique of the semantic representation of Katz & Postal (1964) which he calls "Markerese":

> *we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true. Semantics with no treatment of truth conditions is not semantics (Lewis, 1972, p. 169).*

That is, semantics is not complete without a *model* that provides a way to evaluate logic expressions, in such a way that there is a systematic relationship between different usages of the same entities and predicates[1]. That is, a model for the expressions in Kirby (2002a), as discussed above, needs to recognise that the Peter that *believes* is the same person as the person that *is known to hate*. In predicate logic we can of course define a model that treats *believesthatgavinknowsthatpetehatesheather(x)* as a single predicate that might be true for Peter. But such a model would not do justice to the intended structure of the expression, that is, it would not capture the relation between the statements *"Pete hates Heather"* and *"Gavin knows that Pete hates Heather"*.

For such constructions, **modal logic** provides a more satisfactory framework. A modal logic postulates a set of "possible worlds", each of which has its own set of facts and rules of inference (expressible as, for instance, a predicate logic). For instance, yesterday, today and tomorrow can be seen as three possible worlds. The statement *"it rains"* can have different truth values in each of them. Some statements, however, are true for all worlds (such as *"if it rains, the streets are wet"*). A modal logic, as a minimum, contains the modal operators for necessity '$\Box$' and possibility '$\Diamond$', and a model of the possible worlds. If R denotes *"it rains"*, and W *"the streets are wet"*, then the

[1] In fact, many semanticists now go further and argue that semantics is not complete without a *procedure* to evaluate the expression (Woods, 1968).

statement $\Box(R \mapsto W)$ says *"if it rains, the streets are always wet"*. That statement is true today, if $R \mapsto W$ is true in all worlds that the possible world model gives us access to (in this case, all other worlds, i.e. yesterday and tomorrow).

In particular, a variant of modal logic termed **intensional predicate logic** (IPL) allows us to model the meaning of expressions that describe the content of thoughts, assertions, statements etc. In such a treatment, the beliefs of a person form a possible world, with its own truths. Many problems remain, both internal to IPL (such as the so-called problem of omniscience) and with its relation to natural language constructs (such as so-called donkey-conditionals). For the present purposes, however, it is sufficient to note that there in this hierarchy of logics, more powerful representations like IPL are able to model relations between statements that less powerful representations cannot.

It would be attractive if we could relate these different logics to the assumed differences between the conceptual structures available for communication to modern humans, and those available to their prelinguistic ancestors and modern higher primates. For instance, Jackendoff joins other cognitive scientists in claiming that symbol use is the first major innovation. He does however make clear that he does not believe apes have no symbolic thought:

> *I take it as established by decades of primate research [references omitted] that chimpanzees have a combinatorial system of conceptual structure in place [...]. (Jackendoff, 2002, p. 238)*

The crucial difference, for Jackendoff, is the use of symbolic vocalisations:

> *[Even] single-symbol utterances in young children go beyond primate calls in important respects that are crucial in the evolution of language. Perhaps the most important difference is the non-situation-specificity of human words. The word kitty may be uttered by a baby to draw attention to a cat, to inquire about the whereabouts of a cat, to summon the cat, to remark that something resembles a cat, and so forth. Other primates' calls do not have this property. A food call is used when food is discovered (or imminently anticipated) but not to suggest that food be sought. A leopard alarm call can report the sighting of a leopard, but cannot be used to ask if anyone has seen a leopard lately [references omitted]. (Jackendoff, 2002, p. 239)*

Can we express this intuitive difference between humans and other primates as a difference in representational capacity similar to the difference between propositional and predicate logic? We can, of course, conjecture that humans have words for predicates and other words for objects (the arguments of those predicates), which can thus be used in all situations where the conceptual

system uses that predicate or that object. Primates, on the other hand, can only vocalise complete propositions, even though they, as Jackendoff states, do have a "combinatorial system of conceptual structure".

The problem with such a proposal is that it is not clear a priori what constitutes a predicate or an object, and thus what constitutes situation-specificity in formal terms. How can we be sure that the word *kitty* in an infant's one-word stage does not mean a complete proposition such as *"There is a kitty involved"*? How do we know the child does not simply categorise situations as those that involve kitties, and those that do not, much like a monkey that categorises situations as those that require running into a tree and those that do not? If so, the difference is categorisation, not representational ability. The fact that two different animal species – with different anatomy, and evolved for different habitats – *categorise* the world differently is no surprise, of course.

Conversely, how can we be sure that the meaning of a primate alarm call for leopards is not the predicate *"being a leopard"*? The point here is that with regard to "meanings available for communication" the difference between propositional and predicate logic only shows itself through the rules of combination, that is, through the generalisations they allow. Of course, it is likely that there is something special about the way humans categorise their environment which was crucial for the evolution of language. But the tools of formal semantics do not appear to be useful for characterising that difference.

That leaves us with the conclusion that in terms of formal semantics, use of symbols, Jackendoff's first innovation cannot be recognised (or perhaps even exist) independent from the fourth innovation (concatenation of symbols). In this thesis, I will therefore not address this transition. Perhaps the distinction between predicate and modal logic will prove more useful for characterising the difference between human and non-human thought. A debate exists about whether great apes have thoughts about the thoughts of others, that is, have a theory of mind (Heyes, 1998). Such embedded meanings cannot be modelled with predicate logic. An interesting question is whether the ability for embedded meanings (*I think that she heard that he said...*) is a prerequisite for hierarchical phrase-structure (Dunbar, 1998; Worden, 1998). However, because it plays no role in Jackendoff's scenario, this question will not be addressed in this thesis either.

3.4 Modelling Sound

The mechanisms of sound production and perception in primate and human communication are fortunately more amenable to empirical observation, and there is therefore more of a consensus about the fundamental principles. Sounds are patterns of vibrations, travelling through air from a source (i.e. the speaker) to the hearer. A sound can be represented with a graph that shows the

movements of a membrane (i.e. in a microphone), and hence the variation in sound pressure over time. This graph is referred to as the **waveform**.

For some artificially generated sounds, such as the tone generated by a tuning-fork, the wave-form is a simple, easily interpreted, periodic pattern. For natural calls and utterances, however, the waveform is typically a complex mesh of aperiodic and periodic vibrations of many different frequencies. The analysis of these complex sounds is greatly facilitated by a technique called Fourier Analysis. In Fourier Analysis, the observed waveform is decomposed into (infinitely) many sine-waves of different frequencies, each with a particular amplitude, such that when all these sine-waves are added together the original signal is recovered. A graph that shows for a range of frequencies the amplitude of the corresponding sine-waves is called the **frequency spectrum**.

For both the production and the perception of sounds, the frequency spectrum has a natural interpretation. **Sound production**, both in humans (as worked out by Johannes Mueller in 19th century; see Coren *et al.*, 1979/1994) and many other mammals (Hauser & Fitch, 2003) can be seen as a two-staged process with a vibration source and subsequent filtering (the source-filter model, Chiba & Kajiyama, 1958; Fant, 1960; Titze, 1994). The vibrations are produced by the air flow from the lungs passing the larynx. This sound then propagates through the throat, mouth and nose (the vocal tract), where specific frequencies are reinforced through resonance. The dominant frequency of the source is called the *fundamental frequency*, while resonance frequencies are termed *formants*. Formants can be modified by changing the shape of the vocal tract (opening and closing the jaw, moving the tongue, etc.), thus creating the elementary sounds of a language or call system. In the frequency spectrum, formants show up as distinct peaks in the distribution.

The frequency spectrum also maps in an important way on **sound perception**. When a sound-wave reaches the ear, it sets in motion a cascade of vibrations of the eardrum, hammer, anvil & stirrup, oval window and finally the endolymph fluid in the cochlea. These vibrations cause mechanical waves to travel through the cochlea's fluid. Because of the special shape of the cochlea, the travelling waves reach their maxima at different places along the cochlea's membrane (the "basilar membrane") for each different sound frequency (von Bekesy, 1960; Coren *et al.*, 1979/1994). These differences in wave-form are then translated into different neural activation patterns in the organ of Corti. This way, the mammalian auditory system decomposes an incoming sound-wave in its component frequencies, not unlike the Fourier Analysis performed by phoneticians.

The frequency spectrum is thus a representation of speech sounds that is meaningful for analysing both production and perception. However, the frequency spectrum representation abstracts out the time dimension. Temporal changes in the frequency distribution are crucial for encoding and decoding information into sound in both human and animal communication. A variety of representations of such changes have been developed, including cascade diagrams (where frequency distributions are measured in a number of intervals, and plotted with a small vertical transposition for every subsequent interval) and spectrograms (where frequencies above a specific threshold value are plotted against time). In chapter 4 I will represent such temporal changes as a (discretised) trajectory through an acoustic space (where the coordinates of each point on the trajectory are for instance the peaks in the frequency distribution – the formants – of each small time interval in the waveform).

On the articulatory side, changes in the frequency distribution correspond to movements of articulators in the vocal tract. For example, by abruptly closing the vocal tract and releasing again, an existing sonorous sound is interrupted and a burst of noise is produced (containing many frequencies at low amplitude). In human speech, such movements are perceived as stop consonants (plosives): /p,b,t,d,k,g/, depending on where in the vocal tract the stream of air is interrupted. Similarly, fricatives (e.g. /f,v,s,z/), approximants (/l,r,j,w,h/) and nasal consonants (e.g. /m,n/) are produced by complete or partial blockage, and quick or delayed release. Finally, diphtongs (such as /eɪ/ in *bait*) are produced by a shift in the harmonic quality of the vocal tract without an intermediate consonant.

From a comparative perspective, the basic principles of sound perception and production (at least at the level of physiology of articulators) appear to be very similar across humans and other mammals. In contradiction to the "speech is special" hypothesis (Liberman *et al.*, 1967), recent evidence points at the conclusion that human speech perception is not fundamentally different from non-speech and non-human perception (Hauser & Fitch, 2003; Hauser, 2001). The fact that humans are extra-ordinarily good at perceiving speech sounds appears to be better explained by the observation that, unlike many animal communication systems, human language phonology is *learnt* and *imitated* (Nottebohm, 1976; Studdert-Kennedy, 1983); in the cultural transmission from one generation to the next, languages themselves have evolved to exploit the peaks in performance of the human auditory (and articulatory) systems. I will come back to this point in chapter 7.

However, when analysing the temporal structure of a repertoire of signals, a crucial difference between human and non-human primate communication is noted: human speech is combinato-

rial, that is, the basic meaningful units in human language (words, morphemes) can be analysed as combinations of segments from a small set of basic speech sounds. Semantic animal communication (in the sense of Seyfarth *et al.*, 1980), in contrast, seems to be holistic, that is, the basic meaningful units (calls) cannot generally be decomposed in segments that are reused in other calls. There is some controversy about what the basic segments of human speech are (phonemes, syllables or articulatory gestures), and there are many examples of combinatorial songs in primates, birds, and cetaceans (that are used as sexual display, or for identification; see chapter 4). To my knowledge, no quantitative comparison of the degree of re-use in human and non-human sounds exists. Nevertheless, the intuition that human language exploits this mechanism to an unparallelled degree is widely shared and uncontroversial. This is the third innovation in Jackendoff's list.

At this point it is important to make a distinction between "E-language", the *externally* observable utterances, and "I-language", the system *internal* to the language user that underlies the E-language. Combinatoriality in the I-language can be characterised by defining the basic units and the rules of combination; combinatoriality in the E-language is best characterised by giving a combinatorial I-language that could underlie it. However, the fact that an outside observer can analyse a set of signals as combinatorial, does not necessarily imply that the language user actually exploits that combinatorial potential. Hence, a repertoire of sounds might *superficially* look combinatorial, but in fact not be *productively* combinatorial[2]. For instance, if we accept that the syllable, and not the phoneme, is the unit of productive combination in human speech, then the I-language is characterised by a set of syllables and the rules of combining them. Phonemes, in such a view, are patterns in the E-language that look as if there is a combinatorial system underlying it; they are only superficially combinatorial. This distinction is relevant for the evolution of language, because a superficially combinatorial stage might precede and facilitate the evolution of productive combination (as will be explored in chapter 4 and 5).

Note that for characterising the difference combinatorial and holistic phonology, the basic tools and formalisms from linguistic phonology do not seem to be of much use. In chapter 4 I will describe signals as continuous trajectories through an abstract acoustic space (borrowing a concept from phonetics). In this representation I can model both holistic and combinatorial signals; for the distinction I will rely on visual inspection of the corresponding graphs.

[2]Jackendoff (2002), as do many other linguists, makes the same type of distinction between *productive* and *semi-productive* morphology, but he does not generalise this distinction to the other combinatorial systems in language, nor does he discuss its relevance for evolution.

The evolutionary origins of combinatorial phonology are still a largely open question. A widely shared intuition is that the way to encode a maximum of information in a given time-frame such that it can be reliably recovered under noisy conditions, is by means of a digital code. Hence, phonemic coding could be the result of selection for perceptual distinctiveness. However, this argument has, to my knowledge, never been worked out decisively for human speech (see chapter 4 for a critique of existing formal models, such as the one of Nowak & Krakauer, 1999, and for an alternative proposal).

Alternatively, combinatorial coding could be the result of articulatory constraints. Studdert-Kennedy (1998, 2000) has argued that speech sounds are in fact difficult to produce, and that there is a hierarchy of difficulty of producing speech sounds. This hierarchy is revealed in development. For instance, children master syllables like *"ba"* much earlier than syllables like *"through"*. The reason is that *through* requires a large number of carefully coordinated articulatory movements (gestures). Studdert-Kennedy speculates that the ability to produce such complex sounds is a relatively recent evolutionary innovation, and that the inherent difficulty makes the re-use of motor programs unavoidable. Hence, the combinatorial nature of speech follows from the difficulty of production and the large repertoire of words in human languages.

Consistent with Studdert-Kennedy's scenario is the neurological evidence discussed by Deacon (1997, 2000) that he believes shows intense selection for "the coupling of precisely timed phonation with rapid articulatory movements of tongue, lips and jaw.":

> *Speech, and particularly singing abilities, clearly demonstrate unprecedented forebrain control of the human larynx. In this regard, we are not just divergent from other mammals but also from all other vertebrates — perhaps the only one with significant forebrain control of laryngeal muscles. This is evidence of prolonged intense selection favoring increased vocal abilities in our ancestors. (Deacon, 2000, p.283)*

If Studdert-Kennedy and Deacon are right, Jackendoff's third innovation is characterised by radical changes in articulatory motor control. Nevertheless, this innovation is driven by the need for a large repertoire of perceptually distinct signals, albeit under stringent articulatory constraints. It is possible, as Studdert-Kennedy suggests, that the articulatory constraints already impose a form of combinatorial phonology. In chapter 4, however, I will not make that assumption and instead study its evolution as the result of selection for perceptual distinctiveness alone.

3.5 Modelling Simple Sound–Meaning Mappings

The other transitions in Jackendoff's scenario (nrs. 2 and 4-10 in figure 3.1) all concern the way meanings are mapped on signals. Most existing formalisms in linguistics already assume the innovations that Jackendoff proposes: word order, compositionality, phrase-structure, grammatical categories. They are therefore not of much use in characterising the early transitions. Here I will first develop a simple formalism that describes meaning to form mappings without any assumptions or learning or combination; from that basis I will try to characterise the innovations proposed.

Given a set of relevant meanings to express, and a set of distinctive signals (i.e sounds, or "forms") that can be produced and perceived, we can describe a communication system as a (probabilistic) mapping from meanings to signals (in production), and from signals to meanings (in interpretation). These mappings can be represented with matrices. Hence, we have a *production matrix* $\mathbf{S}$ and an *interpretation matrix* $\mathbf{R}$. $\mathbf{S}$ gives for every meaning m and every signal f, the probability that the individual chooses f to convey m. Conversely, $\mathbf{R}$ gives for every signal f and meaning m, the probability that f will be interpreted as m. If there are M different meanings and F different signals, then $\mathbf{S}$ is a $M \times F$ matrix, and $\mathbf{R}$ a $F \times M$ matrix. Variants of this notation are used by Hurford (1989); Oliphant & Batali (1996) and other researchers.

Many different $\mathbf{S}$ and $\mathbf{R}$ matrices are possible. How can we measure the quality of specific combinations? Or, in biological terms, how can we calculate the *payoff* (a fitness contribution) of specific $\mathbf{S}$ and $\mathbf{R}$ matrices? An important component of such a payoff function is whether speakers and hearers agree on which signals have which meanings. However, in many cases the similarities between signals also need to be taken into account (because more similar signals are more easily confused), as well as the similarities between meanings (because slightly wrong interpretations are often better than totally wrong ones).

The consequences of such similarities can be modelled with a *confusion matrix* $\mathbf{U}$ (of dimension $F \times F$), which gives for each possible signal the probability that it is perceived correctly or as any of the other signals, and with a *value matrix* $\mathbf{V}$ (of dimension $M \times M$), which gives for every intended meaning the *payoff* of each of the possible interpretations. Typically, $\mathbf{U}$ and $\mathbf{V}$ will have relatively high values on the diagonal (the correct signals and interpretations).

Together, these four matrices can describe the most important aspects of a communication system: which signals are used for which meanings by hearers and by speakers, how likely it is that signals get confused in the transmission, and what the consequences of a particular successful

or unsuccessful interpretation are. This notation is a generalisation of the notation in Nowak & Krakauer (1999), and was introduced in Zuidema & Westermann (2003, see appendix C).

A hypothetical example, loosely based on the celebrated study of vervet monkey alarm calls (Seyfarth *et al.*, 1980; Seyfarth & Cheney, 1997), will make the use of this formalism clear[3]. Imagine an alarm call system of a monkey species for three different types of predators: from the air (eagles), from the ground (leopards) and from the trees (snakes). Imagine further that the monkeys are capable of producing a number (say 5) of different signals that range on one axis (e.g. pitch, from high to low) and that these are more easily confused if they are closer together. Thus, the confusion matrix $\mathbf{U}$ might look like the left matrix in figure 3.2.

$U =$

	received signal				
sent signal ↓	16*kHz*	8*kHz*	4*kHz*	2*kHz*	1*kHz*
16*kHz*	0.7	0.2	0.1	0.0	0.0
8*kHz*	0.2	0.6	0.2	0.0	0.0
4*kHz*	0.0	0.2	0.6	0.2	0.0
2*kHz*	0.0	0.0	0.2	0.6	0.2
1*kHz*	0.0	0.0	0.1	0.2	0.7

$V =$

	interpretations		
intentions ↓	*eagle*	*snake*	*leopard*
eagle	0.9	0.2	0.1
snake	0.5	0.9	0.5
leopard	0.1	0.2	0.9

Figure 3.2: Confusion and value matrices for the monkeys in the example, describing the noise in signalling and the value of intention–interpretation pairs in their environment.

Further, although it is obviously best to interpret a signal correctly, if one makes a mistake, typically not every mistake is equally bad. For example, if a leopard alarm is given, the leopard response (run into a tree) is best, but a snake response (search surrounding area) is better than an eagle response (run into a bush, where leopards typically hide) (Seyfarth & Cheney, 1997). Thus the value matrix $\mathbf{V}$ might look somewhat like the right matrix in figure 3.2.

$S =$

	sent signal				
intention ↓	16*kHz*	8*kHz*	4*kHz*	2*kHz*	1*kHz*
eagle	1.0	0.0	0.0	0.0	0.0
snake	0.0	0.0	1.0	0.0	0.0
leopard	0.0	0.0	0.0	0.0	1.0

$R =$

	interpretation		
received signal ↓	*eagle*	*snake*	*leopard*
16*kHz*	1.0	0.0	0.0
8*kHz*	1.0	0.0	0.0
4*kHz*	0.0	1.0	0.0
2*kHz*	0.0	0.0	1.0
1*kHz*	0.0	0.0	1.0

Figure 3.3: Production and interpretation matrices for the monkeys in the example, describing which signals they use for which meanings.

Now, assume a speaker i with her $\mathbf{S}^i$ as the left matrix in fig. 3.3, and a hearer j with his $\mathbf{R}^j$ as the right matrix in that figure. What will happen in the communication between i and j? One possibility is that (i) the speaker sees an eagle, (ii) she sends out the 16*kHz* signal, (iii) the hearer

[3]The actual alarm calls of vervet monkeys are very different from the ones I use in this example. For instance, eagle alarm calls are low-pitched rather than high-pitched, and all three types of alarm calls have a temporal structure. The example here is just to illustrate the use of the formalism.

indeed perceives this as a $16kHz$ signal and (iv) he correctly interprets this signal as "eagle". The contribution to the expected payoff is:

$$\begin{aligned} P_{\mathbf{S}}(16kHz\ sent \mid eagle\ observed) &\times \\ P_{\mathbf{U}}(16kHz\ perceived \mid 16kHz\ sent) &\times \\ P_{\mathbf{R}}(eagle\ interpreted \mid 16kHz\ perceived) &\times \\ \mathbf{V}(eagle\ interpreted, eagle\ observed)) &= 1 \times .7 \times 1 \times .9 = 0.63. \end{aligned} \tag{3.1}$$

Another possibility, with probability 0.2, is that the hearer misperceives the signal as a $8kHz$ signal, but with probability 1 still interprets it correctly. We can thus work out all possible scenarios and find that the expected payoff w_{ij} of the interaction between i and j, given the constraints on communications as in $\mathbf{U}$ and $\mathbf{V}$ in fig. 3.2, is: $w_{ij} = .7 \times .9 + .2 \times .9 + .1 \times .2 + .2 \times .5 + .6 \times .9 + .2 \times .5 + .1 \times .2 + .2 \times .9 + .7 \times .9 = 2.4$.

More generally, such calculation can be expressed by one simple expression for the expected payoff w_{ij} of communication between a speaker i with production matrix $\mathbf{S}^i$ and a hearer j with interpretation matrix $\mathbf{R}^{\mathrm{j}}$ (Zuidema & Westermann, 2003):

$$w_{ij} = \mathbf{V} \cdot \left(\mathbf{S}^i \times \left(\mathbf{U} \times \mathbf{R}^j\right)\right). \tag{3.2}$$

In this formula, "$\times$" represents the usual matrix multiplication and "$\cdot$" represents dot-multiplication (the sum of all multiplications of corresponding elements in both matrices; the result of dot-multiplication is not a matrix, but a scalar).

In this simple example, the matrices $\mathbf{U}$ and $\mathbf{V}$ are very small, and reflect only a 1-dimensional topology in both signal and meaning space. The matrices $\mathbf{S}$ and $\mathbf{R}$ are set by hand to arbitrarily chosen values. In contrast, in the simulations of chapter 5 I will consider larger and more complex choices for $\mathbf{U}$ and $\mathbf{V}$, and I will use a hill-climbing algorithm to find the appropriate (near-) optimal settings for $\mathbf{S}$ and $\mathbf{R}$.

Note that the $\mathbf{S}$ and $\mathbf{R}$ matrix describe the production and interpretation behaviour of an individual (the E-language), but they do not necessarily model the mechanism that individuals use to map meaning onto signals and vice versa (the I-language). The values can even be chosen such that an individual's $\mathbf{S}$ matrix is incompatible with her own $\mathbf{R}$ matrix, that is, that she cannot understand her own utterances. More realistic, perhaps, is to assume an underlying lexicon of (bi-directional) associations between meanings and signals (Steels, 1995; Komarova & Niyogi,

2004). Such associations can be modelled with an association matrix **A**. **S** and **R** are then functions of **A**, such that for instance, an element in **S** is 1 if the corresponding element in **A** is the highest in its row, and 0 otherwise. Similarly, an element **R** can be set to 1 only if the corresponding element in **A** is the highest in its column.

Jackendoff's second innovation – an open, unlimited class of symbols – can be viewed as the evolution of a learning procedure to set the values of the elements in **S** and **R** or **A**. Assume that the set of possible relevant meanings and the set of possible signals are determined by an individual's habitat and anatomy and can be defined a priori. An innate, closed call system then corresponds to settings of the elements of the matrices that are not dependent on input and show no variation; conversely, a learnt, open call system corresponds to settings that do depend on environmental input and vary with varying inputs. Innate calls appear to be the norm in primate communication. For instance, Seyfarth & Cheney (1997) argue that, although there might be social learning of response behaviour (interpretation), the production of calls in most primates must be considered innate.

Human language, in contrast, clearly is an open system, where the meanings of words are not naturally given, but rather emerge as conventions in a population of language users (Lewis, 1969; Gamut, 1991). Conventions are *negotiated* in a population, by individuals learning from (and adapting to) each other. Different learning strategies lead to different languages, and have different consequences for the biological fitness of individuals in a population, as is studied by (Hurford, 1989) and others. The main results from these studies is that (i) a stable communication system can emerge in a population where everybody learns from everybody else, without a need for central control (Steels, 1995); (ii) the best "response language" is not necessarily the same as the current language in the population (Hurford, 1989; Komarova & Niyogi, 2004), and (iii) Saussurean learners (where S matrices are modelled after R matrices) and synonymy and homonymy avoiders outcompete other learning strategies (Hurford, 1989; Oliphant & Batali, 1996; Smith, 2004).

These studies are interesting, but do not really address Jackendoff's transition from a closed, innate vocabulary to an open, learnt vocabulary. The selective advantages of such a transition – to what biologists call "phenotypic plasticity" – depend on the constraints on the innate system, the properties of the environment and the accuracy of the learning mechanism. If a learnt vocabulary can contain more signals than an innate vocabulary – as Jackendoff asserts – that must be because of biological constraints preventing the innate system to be as large. Moreover, a learnt vocabulary can be easily extended to include a word for a new concept, but whether or not this confers

an advantage depends on how often such new relevant concepts appear. These are interesting issues, but it is difficult to tell what reasonable assumptions are. Oliphant (1999) argues quite convincingly that the computational demands of learning are unlikely to have been the limiting factor in this transition; rather, he argues, the difficulties to identify what meaning a signal is meant to convey explain why learnt communication systems are so rare in nature.

In conclusion, I agree with Jackendoff (2002), Oliphant (1999) and many others that the emergence of an open class of symbols is an important transition in the evolution of language. Moreover, I believe it can be formalised using the matrix notation introduced above. Many models that use such a formulation in one form or another have been studied. In this thesis, however, I will not address this problem. In chapter 5 I will study a model where I simply assume that every learning step will increase an agent's ability to communicate with others (i.e. they optimise their $\mathbf{S}$ and $\mathbf{R}$ matrices), and focus on the effects of different $\mathbf{U}$ and $\mathbf{V}$ matrices. As we will see, specific choices for these matrices have consequences for the evolution of both combinatorial phonology and compositional semantics.

3.6 Modelling Compositionality

The matrices discussed above can describe, for each particular meaning, which signals are associated with it or vice versa. However, they cannot make explicit any *regularity* in the mapping from meanings to signals. In both non-human and human communication such systematic relations between meanings and signals exist. For instance, in most species high pitch sounds are associated with danger, and low pitch sounds with aggression. In vervet-monkey calls, there are clear similarities between the calls used in social interactions, which are all very different from the alarm calls (Seyfarth & Cheney, 1997). In human language, on the level of words, there is some evidence – albeit controversial – that similar words tend to refer to similar objects, actions or situations, and that humans generalise such patterns to nonsense words (Hinton *et al.*, 1995). Uncontroversially, on the level of morphosyntax it is clear that similar sentences mean similar things, that is, the mapping from meanings to signals is compositional.

We can describe the systematicities in the meaning–signal mappings as the preservation of topology between meaning space and signal space, that is, meanings that are close are expressed with signals that are close. In the $\mathbf{S}$, $\mathbf{R}$ and $\mathbf{A}$ matrices, such "topology preservation" might be noticeable as a non-random pattern if both the meaning and signals axes are ordered. We can be more precise, however, by systematically comparing each pair of associations. Brighton

(2002) proposes using the correlation ("Pearson's r") between the distance between each pair of meanings and the distance between the corresponding signals:

$$r = \underset{m,m' \in M}{\text{correlation}} \left(D\left(m, m'\right), D\left(S[m], S[m']\right)\right), \tag{3.3}$$

where $S[m]$ gives the most likely signal used to express m according to $\mathbf{S}$, $D(m, m')$ gives the distance between two meanings m and m', and $D(f, f')$ between two signals f and f'. Although only a correlate of compositionality, such a measure can reveal a tendency for related meanings to be expressed with related signals. Hence, expressed in this formalism, Jackendoff's fourth and fifth innovation (concatenation and compositionality) correspond to a high values of r in equation (3.3).

We can go further, however, and explicitly model the way in which combinations of signs form more complex signs. The common way to deal systematically with the meanings of *combinations* of lexical entries, is Church's lambda calculus (see e.g. Gamut, 1991, for a discussion). Semantic descriptions, such as discussed in section 3.3 should be extended with the possibility to include lambda (λ) terms. Lambda terms can be seen as listing the variables that still need to be substituted; they disappear when a complete semantic description is reached. Formally, a lambda term in front of an expression turns that expression into a function that maps an argument onto a new expression where that argument has found its proper place. For instance, we can model the semantics of the verb *walks* as $\lambda x W(x)$ and apply it to an argument j (for *John*) to yield $W(j)$ (for *John walks*). Similarly, the following is the semantic description for *approaches* in the "flat notation" of De Beule *et al.* (2002):

$$\lambda x \lambda y (x \mid (approach\ z)\ (agent\ z\ x)\ (patient\ z\ y)) \tag{3.4}$$

When applied to the following semantic description

$$p \mid (circle\ p) \tag{3.5}$$

the resulting description is as follows:

$$\lambda y (p \mid (approach\ z)\ (agent\ z\ p)\ (patient\ z\ y)\ (circle\ p)) \tag{3.6}$$

That is, the variable x in (3.4) is replaced by the head of (3.5), and the λx is removed. (3.6) means something like "the circle approaches y", where y still needs to be filled in. (3.6) can in

turn be applied to, for instance, the description of a block, yielding a description that would mean *"The circle approaches the block"*.

The lambda calculus gives a mechanical procedure to derive the semantic expression that results from applying a function to an argument. A word (or phrase) corresponding to the function is said to *dominate* a word corresponding to the argument. Hence, if we model the compositional semantics of *"John walks"* with a function $\lambda x W(x)$ and an argument j, then we have assumed that *walks* dominates *John*.

In modern languages, this dominance structure is, to a large extent, determined by principles of word order and morphological marking. Thus, if we model the semantics of the transitive verb *hates* in *"George hates broccoli"* as $\lambda y \lambda x H(x, y)$ (i.e. as a function with two arguments), the principles of word order need to guarantee that *hates* dominates *broccoli*, and *hates broccoli* dominates *George*. In the fourth and fifth stage of Jackendoff's scenario there is no phrase-structure or morphological marking, so the dominance structure is largely underdetermined. The word order principles of "agent first", "focus last" and "grouping" that Jackendoff proposes, constrain the structural ambiguity that arises from this underdeterminacy.

In conclusion, we have with equation (3.3) a provisional measure of compositionality in the E-language. Moreover, we can characterise compositionality in the I-language by identifying the units and rules of combination. In chapter 5 I will study the transition to compositional semantics using the former, and argue that the compositionality in the I-language can more easily evolve if some form of compositionality in the E-language has already been established for other reasons.

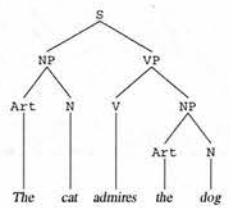
3.7 Modelling Hierarchical Phrase Structure

One of the defining characteristics of human language is that sentences exhibit phrase-structure, and the ability to represent phrase-structure has, since Chomsky (1957), been the one of the most important criteria in judging the adequacy of linguistic formalisms. Rewriting grammars, such as proposed by Chomsky, remain the archetype formalism for describing syntax. I will first introduce this formalism in some detail, and then define phrase-structure in terms of it.

Rewriting grammars are specified by four sets: the terminal symbols, the nonterminal symbols, the production rules and the start symbols. Terminal symbols (V_t) are in a sense the atoms of a language without further syntactic structure (e.g., the words or morphemes of a language) but possibly also complete idioms or frozen expressions. Nonterminal symbols (V_{nt}) are variables that stand for constituents of a sentence, and can correspond to anything from the syntactic category of a word (or morpheme) to a whole sentence. Production rules (R) specify which nonterminal symbols can be replaced by which terminal or non-terminal symbols in the process

S	$\mapsto$	NP VP	(1)	Art	$\mapsto$	*the*	| *a*	(6*ab*)
NP	$\mapsto$	Art N	(2)	N	$\mapsto$	*cat*	| *dog*	(7*ab*)
N	$\mapsto$	N SP	(3)	V	$\mapsto$	*chases*	| *admires*	(8*ab*)
VP	$\mapsto$	V NP	(4)	S	$\mapsto$	*the cat fears the dog*		(9)
SP	$\mapsto$	*that* VP	(5)	S	$\mapsto$	*the dog fears the cat*		(10)

(a)



(b)

Figure 3.4: An example context-free grammar for a fragment of English, with a terminal alphabet $V_t = \{$*the, a, dog, chases, admires, that*$\}$ and a non-terminal alphabet $V_{nt} = \{\mathtt{S}, \mathtt{NP}, \mathtt{VP}, \mathtt{Art}, \mathtt{N}, \mathtt{V}\}$. Production of a sentence ("derivation") always starts with the symbol S and proceeds by replacing symbols matching the left-hand side of some rule, with the string at the right-hand side of that rule. "Parsing" means searching the sequence of rewriting steps that would produce the sentence with a given grammar. Rules 1–4 are "combinatorial"; rule 5 is "recursive". The grammar can generate infinitely many sentences such as "the cat chases the dog" or "a dog admires a cat that chases a dog that admires a cat" etc. Rule 6–8 constitute what is traditionally described as the lexicon, and can be represented in the same formalism. Rule 9 and 10 illustrate a "lexical", non-combinatorial and much less efficient strategy for generating sentences. Context-free grammars are considered to be not quite sufficiently powerful to describe natural languages. The formalism can be extended in several ways. For instance, it can be extended to attribute in a systematic way meanings to words and sentences; the resulting system is "compositional".

of deriving a sentence, starting with a start-symbol (S). If the production rules are of the form $\alpha \mapsto \omega$ where the lefthand-side is a nonterminal symbol ($\alpha \in V_{nt}$), and the righthand-side is any (non-null) string of terminals and nonterminals, the grammar is said to be context-free (because the context in which α occurs is not relevant for the applicability of the rule). Figure 3.4 gives an example context-free rewriting grammar for a fragment of English.

Chomsky (1957) showed that more restricted versions of this formalism such as finite-state grammars or their probabilistic version, Markov processes, are unable to describe the long-range dependencies that are observed in natural languages. English, for instance, requires agreement between the number of a subject and the number of the verb, such as in example 3.7(a), but not 3.7(b). But English also allows the insertion of one or more prepositional phrases such as "that

chases the dog" as a modifier of the subject, such as in examples 3.7(c, d). Hence, the distance between "the cat" and "admires" can –in principle– be arbitrarily long, and the dependency between them cannot be described in terms of transition probabilities between words.

(3.7) a. The cat admires the dog.

b. *The cat admire the dog[4].

c. The cat that chases the dog admires the dog.

d. The cat that chases the dog that admires the cat admires the dog.

Crucially, adequate formalisms for natural language need to represent *phrase-structure*, that is, they need to recognise that "the cat", "the cat that chases the dog" and "the cat that chases the dog that admires the cat" all have the same role in their respective sentences. Moreover, phrase-structure is *hierarchical*; the first "the dog" in example 3.7(c) is a noun phrase that is a component of the whole subject noun phrase. The hierarchical phrase-structure of a sentence can be visualised as a tree, as is exemplified in figure 3.4(b). Phrase-structure can be viewed as the necessary intermediate representation in mapping complex sounds ("phonological form", **PF**, in Chomsky's terminology) to complex meanings ("logical form", **LF**). It specifies both the word order ("linear precedence") and the dominance structure ("immediate dominance").

Context-free grammars can represent hierarchical phrase-structure. Starting with a start symbol, one can iteratively apply the production rules by replacing an occurrence of the left-hand side of a rule by the right-hand side. The phrase-structure of figure 3.4(b) can be *derived* as follows:

$$\mathrm{S} \circ 1 \circ 2 \circ 6a \circ 7a \circ 4 \circ 8b \circ 6a \circ 7b = \textit{"the cat admires the dog"}, \tag{3.8}$$

where $t \circ r$ gives the result of applying rule r to a tree t (I here assume it is applied to the left-most nonterminal in tree t, which ensures that there is a single derivation for each unique phrase-structure tree).

Parsing refers to a search procedure for finding the phrase-structure and corresponding derivation for a given sentence and a given grammar. For the purposes of this chapter it is sufficient to note that parsing is a non-trivial problem. Naive, exhaustive, bottom-up search algorithms have a time complexity that is exponential in the number of words in a sentence; with some clever book-keeping to avoid redundant work, the time complexity can be reduced to $O(n^3)$, which is still a significant limitation for real-world applications and for cognitive realism. Such problems

[4]The asterix indicates, in line with linguistic convention, that this sentence is ungrammatical.

are, in Chomsky's view, part of the domain of *performance*, and not of major concern for theorists of language *competence*.

Context-free grammars are a powerful formalism, but Chomsky (1957) argued (for the wrong reasons, it emerged later) that they are not powerful enough to model certain phenomena in language. Chomsky's examples are the following:

(3.9) a. the scene of the movie was in Chicago
b. the scene of the play was in Chicago
c. the scene of the movie and of the play was in Chicago

(3.10) a. the scene of the movie was in Chicago
b. the scene that I wrote was in Chicago
c. *the scene of the movie and that I wrote was in Chicago

Chomsky argued that the rule to describe the proper use of "and" (*coordination*), as in example 3.9(c) but not 3.10(c), requires itself knowledge of the phrase-structure. That is, the "and" rule is a meta-rule that uses the phrase structure of 3.9(a) and (b) to decide that (c) is possible because "the movie" and "the play" are of the same type, and, conversely, uses the phrase-structure of 3.10(a) and (b) to decide that (c) is not possible. Such meta-rules were termed *transformations* and were used not only for coordination, but for many other common linguistic constructions such as questions (wh-movement), gapping, passives and topicalization.

With his discussion of finite-state grammars, context-free grammars and transformations, Chomsky discovered a fundamental hierarchy of grammars that is now termed the Chomsky Hierarchy (see table 3.1). This prompted a long (and still continuing) debate on where to locate human language on the hierarchy. Already in the 1960s it was realised that the original transformational grammars, which are context-sensitive, were too powerful, because they made necessary a long list of rather ad-hoc constraints and exceptions to exclude obviously ungrammatical sentences (for instance, the "coordinate structure" constraint and the "across the board exception" in Ross, 1967). In the 1970s and 1980s efforts were made to find a formalism that could do away with tranformations, but would still be sufficiently powerful to deal with phenomena like coordination, wh-movement and gapping. Solutions, such as GPSG, essentially work by systematically increasing the number of non-terminals enormously (that is, they give with slashes or indices the non-terminals an internal structure).

There seems to be a relative consensus now that the necessary level of generative power is slightly more than context-free, a level now termed "mildly context-sensitive" (Joshi *et al.*, 1991). The additional power over context-free grammars is needed for relatively rare constructions such

Definition 1 ***(Chomsky hierarchy)*** *A grammar* $G = \langle P, S, V_{nt}, V_{te} \rangle$ *is classified according to the following restrictions on the form of rewriting rules of* P*:*

1. *A grammar is of* TYPE 3 *(the "right-linear" or "regular grammars") if every rule is of the form* $A \mapsto bC$ *or* $A \mapsto b$*, where* $A, C \in V_{nt}$*, and* $b \in V_{te}^*$ *or* $b = \lambda$ *(the "empty" character).*
2. *A grammar is of* TYPE 2 *(the "context-free grammars") if every rule is of the form* $A \mapsto w$*, where* $A \in V_{nt}$*, and* $w \in (V_{nt} \cup V_{te})^*$*.*
3. *A grammar is of* TYPE 1 *(the "context-sensitive grammars") if every rule is of the form* $vAw \mapsto vzw$*, where* z *is any combination of terminal or non-terminal symbols:* $z \in (V_{nt} \cup V_{te})^*$ *and* $z \neq \lambda$*. In addition, a single rule* $S \mapsto \lambda$ *is allowed if* S *does not appear at any right-hand side of the rules.*
4. *Any rewriting grammar, without restrictions, is of* TYPE 0*.*

This classification constitutes a strict hierarchy of languages. Hence, if $\mathcal{L}_i$ is the set of all languages of type i, then the following is true:

$$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0. \tag{3.11}$$

Table 3.1: The Chomsky Hierarchy

as the crossed-dependencies in the Dutch example 3.12(a). Examples (b) and (c) show the translation in English and German. Different fonts are used to show the different types of dependencies: crossing in dependencies in Dutch, local dependencies in English, center-embedding in German.

(3.12) a. Gilligan beweert dat **Kelly** *Campbell* BLAIR het publiek **zag** *helpen* BEDRIEGEN.

b. Gilligan claims that **Kelly saw** *Campbell help* BLAIR DECEIVE the public.

c. Gilligan behaupte dass **Kelly** *Campbell* BLAIR das Publikum BELÜGEN *helfen* **sah**.

Given a formal definition of complexity in terms of the Chomsky Hierarchy, and a consensus about where modern human language should be situated, it is perhaps natural to try to describe the evolution of language as a climb of that hierarchy. In such a scenario, selection for increased computational power removed one by one the computational constraints for dealing with the full complexity of modern language. An explicit example of such a scenario is Hashimoto & Ikegami (1996), but it is implicit in many other accounts (e.g. Fitch & Hauser, 2004).

However, there are a number of problems with such an attempt. First, we have to be very careful with what we mean with phrases like "at least context-free power", "human language syntax is mildly-context sensitive" or "where human language is situated on the Chomsky Hierarchy". The classes of formal languages on the Chomsky Hierarchy are subsets of each other

(equation (3.11)). Chomsky's (1957) analysis that finite state grammars are insufficient, and subsequent analysis that context-free grammars are also insufficient, suggests that natural languages are in that subset of the context-sensitive languages that cannot be modelled with a finite-state grammar or context-free grammars (that is, in the *complement* of the context-sensitive in the context-free languages). Most context-sensitive languages, however, have very little in common with natural languages; natural languages are thus constrained in many ways (e.g. semantics, learnability) that have nothing to do with the Chomsky Hierarchy.

Second, it would be wrong to assume that complexity in terms of the Chomsky Hierarchy is actually hard to get. Just a few neurons connected in specific way can generate temporal patterns that are of type 1 or 0 in the Chomsky Hierarchy (i.e. that can only be described with context-sensitive or Turing complete grammars). Like natural language, such patterns would justify the label "at least context-sensitive", even though they are not likely to be interesting from the point of view of encoding information. In short, the classes of the Chomsky Hierarchy divide up the space of formal grammars in a way that is not particularily relevant for the evolution of language. That is, it is possible that most of the evolutionary developments of natural language grammar occurred within one and the same class on the Chomsky Hierarchy. Moreover, even if a class boundary was crossed, formalisation in terms of the Chomsky Hierarchy and architectural constraints offers no insights about the causes for crossing it.

Are there ways to divide up the space of formal grammars that do suggest an incremental, evolutionary path to the complexity of modern language? A starting point for answering that difficult question might be a precise model of how natural language is learnt. In chapter 6 I will discuss a learning algorithm for natural language grammar. I will argue that – before we can usefully propose an incremental pathway for the evolution of syntax (including Jackendoff's remaining innovations of grammatical categories, inflectional morphology and grammatical functions) – we need to recognise the fact that language learning is a peculiar learning problem. Languages are transmitted culturally, and can therefore lead to a form of cultural evolution. As we will see in these chapters, the incremental evolution of the human capacity for language can only be understood as a co-evolution of languages and the brain (Deacon, 1997).

3.8 Conclusions

Humans show in their communication system a number of "design features" that distinguish us from non-human primates and, by assumption, from our pre-linguistic ancestors. Jackendoff's list of innovations in the evolution of languages provides a useful framework to address the origins of these design features. Jackendoff's account can, however, be improved in a number of ways.

First, I have argued that although a scenario with successive stages is an important ingredient of an evolutionary explanation, Jackendoff does not address the important other ingredient: how did the transitions happen? Evolutionary explanations require a plausible account of how innovations spread in the population.

Second, although Jackendoff makes liberal use of diagrams, trees and logic formulae, his account is not precise enough to be implemented in formal models. In this chapter I have tried to sketch the formal tools we need to describe evolutionary innovations in meanings, sounds and the mapping between them. From that discussion it has become clear that Jackendoff's first innovation, the use of symbols, cannot be precisely defined. In contrast, combinatorial phonology, compositional semantics and hierarchical phrase-structure can be precisely characterised. These innovations will be the topics of the next three chapters.

Third, Jackendoff's evolutionary scenario does not make a distinction between the structure of the language as observed from 'outside' (E-language), and the structure of the representations used in an individual's brain (I-language). As we have seen in the discussion of this chapter, it is possible for a language to show the hallmarks of combinatorial phonology, compositional syntax and perhaps phrase-structure, without the language user being able to actively exploit it. In the chapters 4 and 5, I will explore how that observation can be a solution to the problem of coordination discussed in chapter 2.

CHAPTER 4

Combinatorial Phonology[1]

A fundamental, universal property of human language is that its phonology is combinatorial. That is, one can identify a set of basic, distinct units (phonemes, syllables) that can be productively combined in many different ways. In this chapter, I review a number of theories and models that have been developed to explain the evolutionary transition from holistic to combinatorial signal systems, but find that in all problematic linguistic assumptions are made, or crucial components of evolutionary explanations are omitted. I present a novel model to investigate the hypothesis that combinatorial phonology results from optimising signal systems for perceptual distinctiveness. The model differs from previous models in two important respects. First, signals are modelled as trajectories through acoustic space. Hence, both holistic and combinatorial signals have a temporal structure. Second, I use the methodology from evolutionary game theory. Crucially, I show a path of ever increasing fitness from holistic to combinatorial signals, where every innovation represents an advantage even if no-one else in a population has yet obtained it.

[1]This chapter describes research that builds on joint work with Bart de Boer, as published in de Boer & Zuidema, 2003 (see appendix C of this thesis). However, all modelling, text and graphs in this chapter are my own, except where indicated otherwise.

4.1 Introduction

4.1.1 Natural language phonology is combinatorial

One of the universal properties of human language is that its phonology is *combinatorial.* In all human languages, utterances can be split into units that can be recombined into new valid utterances. Although there is some controversy about what exactly the units of (productive) combination are, there is general agreement that in natural languages – including even sign languages (Deuchar, 1996) – meaningless atomic units (phonemes or syllables) are combined into larger wholes; these meaningful combinations (words, or morphemes) are then further combined into meaningful sentences. These two levels of combination constitute the *duality of patterning* (Hockett, 1960).

In the traditional view, the atomic units are *phonemes* (minimal speech sounds that can make a distinction in meaning), or the distinctive features of these phonemes (Chomsky & Halle, 1968). Signal repertoires that are built-up from combinations of phonemes are said to be "phonemically coded" (Lindblom *et al.*, 1984). For instance, the words "we", "me", "why" and "my", as pronounced in standard British English, can be analysed as built-up from the units "w", "m", "e" and "y", which can all be used in many different combinations. One popular alternative view is that the atoms are *syllables*, or the possible onsets, codas and nuclei of syllables (e.g. Levelt & Wheeldon, 1994). A second alternative theory uses *exemplars*, which can comprise several syllables or even words, as its basic units (e.g. Pierrehumbert, 2001). In this chapter, I will avoid the debate about the exact level of combination – and the conventional term "phonemic coding" – and instead focus on the uncontroversial abstract property of "combinatorial phonology"[2].

Note that, whichever the real level of combination is, there is no logical necessity to assume that all recurring sound patterns observed in speech, are in fact units of productive combination in the speaker's brain. For instance, if one accepts that syllables or exemplars are the units of combination used by the speaker, phonemes are still a useful level of description to characterise differences in meaning. I distinguish between:

1. *productively combinatorial phonology*, where the cognitive mechanisms for producing, recognising and remembering signals make use of a limited sets of units that are combined in many different combinations. Productive combinatoriality is a property of the internal representations of language in the speaker (I-language).

[2]In the animal behaviour literature the term "phonological syntax" (coined by Peter Marler, see Ujhelyi, 1996) is often used, and Michael Studdert-Kennedy also uses the term "particulate principle" (coined by W. Abler, see Studdert-Kennedy, 1998). Jackendoff (2002, p.238) uses the term "combinatorial, phonological system" on which my terminology is based.

2. *superficially combinatorial phonology*, where parts of signals overlap with parts of other signals. Superficial combinatoriality is a property of the observable language (E-language). Importantly, the overlapping parts of different signals need not necessarily also be the units of combination of the underlying linguistic representations.

This chapter is concerned with mathematical and computational theories of the evolution of combinatoriality of human languages at both these levels. It has often been observed that natural language phonology is *discrete*, in that it allows only a small number of basic sounds and not all feasible sounds in between. In this chapter, I argue that it is important to distinguish between discreteness per se, and superficial and productive combinatoriality. In section 4.2, I will review existing models of Liljencrants & Lindblom (1972), Lindblom, MacNeilage & Studdert-Kennedy (1984), de Boer (2001) and Oudeyer (2001, 2002), and argue that they are relevant for the origins of discreteness, but have little to say about the origins of superficial and combinatorial phonology. Nowak & Krakauer (1999) do address the origins of productive combinatoriality, but their model has a number of shortcomings that make it unconvincing as an explanation for its evolution.

In my own model, that I will introduce in section 4.3, I address the questions of why natural language phonology is both discrete and superficially combinatorial. I assume, but do not show in this chapter, that superficial combinatoriality is an important intermediate stage in the evolution of productive combinatoriality.

4.1.2 The origins of combinatorial phonology

Although discrete, combinatorial phonology has often been described as a uniquely human trait (e.g. Hockett, 1960; Jackendoff, 2002), it is increasingly realised that many examples of bird and cetacean songs (e.g. Doupe & Kuhl, 1999; Payne & McVay, 1971) and, importantly, non-human primate calls are combinatorial as well (Ujhelyi, 1996). For instance, the "long calls" of tamarin monkeys are built up from many repetitions of the same element (e.g. Masataka, 1987), and those of gibbons (e.g. Mitani & Marler, 1989) and chimpanzees (e.g. Arcadi, 1996) of elaborate combinations of a repertoire of notes.

Such comparative data should be taken seriously, but it is unwarranted to view combinatorial long calls in other primates as an immediate precursor of human combinatorial phonology, because there are some important qualitative differences:

- Although a number of building blocks might be used repeatedly to construct a call, it does not appear to be the case that rearranging the building blocks results in a call with a different meaning.

- It is unclear to what extent the building blocks of primate "long calls" are flexible and whether they are learnt.
- In human language, combinatorial phonology functions as one half of the "duality of patterning": together with recursive, compositional semantics it yields the unlimited productivity of natural language, but it is unclear if the single combinatorial system of primates can be seen as its precursor.

Nevertheless, combinatorial phonology must have evolved from holistic systems by natural selection. There are at least two views on what the advantages of combinatorial coding over holistic coding are:

1. It makes it possible to transmit a larger number of messages over a noisy channel (the "noise robustness argument", an argument from information theory, e.g. Nowak & Krakauer, 1999). Note that this argument requires that the basic elements are distinct from each other, and that signals are strings of these basic elements. The argument does not address, however, how signals are stored and created;
2. It makes it possible to create an infinitely extensible set of signals with a limited number of building blocks. Such productivity provides a solution for memory limitations, because signals can be encoded more efficiently, and for generalisation, because new signals can be created by combining existing building blocks (the "productivity argument", a point often made in the generative linguistics tradition, e.g. Jackendoff, 2002). Note that this argument deals purely with the cognitive aspects, and views the acoustic result more as a side-effect.

These views are a good starting point for investigating the question of *why* initially holistic systems (which seem to be the default for smaller repertoires of calls) would evolve toward combinatorial systems. However, as I explored in chapter 2, just showing an advantage does not constitute an evolutionary explanation. At the very least, evolutionary explanations of an observed phenotype, involve a characterisation of (i) the set of possible phenotypes, (ii) the fitness function over those phenotypes, and (iii) a sequence of intermediate steps from an hypothesised initial state to observed phenotype. For each next step, one needs to establish that (iv) it has selective advantage over the previous, and thus can invade in a population without it. In section 4.2 I will criticise some existing models because they lack some of these required components.

In language evolution, fitness will not be a function of the focal individual's traits alone, but also of those of its conversation partners. That is, the selective advantage of a linguistic trait will depend on the frequency of that trait and other traits in a population (it is "frequency dependent"). Therefore, evolutionary game-theory (Maynard Smith, 1982) is the appropriate

framework for formalising evolutionary explanations for language (Nowak & Krakauer, 1999; Komarova & Nowak, 2003; Smith, 2004; van Rooij, 2004; Jäger, 2005). In this framework, the crucial concept is that of an evolutionary stable strategy (henceforth, ESS): a strategy that cannot be invaded by any other strategy (Maynard Smith & Price, 1973). Thus formulated, the challenge is to show that (i) repertoires of signals with a combinatorial phonology are ESSs, and that (ii) plausible precursor repertoires, without combinatorial phonology, are not evolutionarily stable.

There are also theories of the origins of combinatorial phonology that do not assume a role for natural selection. For instance, Lindblom *et al.* (1984), de Boer (2001) and Oudeyer (2001, 2002) see "self-organisation" as the mechanism responsible for the emergence of combinatorial phonology. These authors use the term self-organisation in a very broad sense, where it can refer to almost any process of pattern formation other than classical, Darwinian evolution. Liljencrants & Lindblom (1972) use an optimisation heuristic, but do not make explicit which process underlies the optimisation. In the next section I will argue that self-organisation and natural selection need not be put in opposition, but can be seen as detailing *proximate* and *ultimate* causes respectively (Tinbergen, 1963; Hauser, 1996), where natural selection modifies the parameters of a self-organising process (Waddington, 1939; Boerlijst & Hogeweg, 1991).

4.2 Existing Approaches

4.2.1 Maximising discriminability

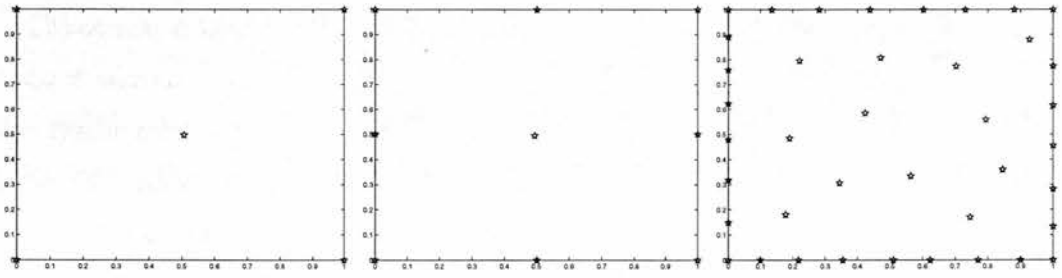
Liljencrants & Lindblom (1972) argued that one can understand the structure of the sound systems in natural language as being optimised for perceptual discriminability and articulatory ease, rather than as arbitrary settings of parameters (as in the theories from the generative phonology tradition, e.g. Chomsky & Halle, 1968). In the initial paper they focused on the discriminability of vowel repertoires, and proposed the following metric to measure their quality:

$$E = \frac{1}{2} \sum_{i,j \neq i \in R} \frac{1}{d_{ij}^2} = \sum_{i=2}^{|R|} \sum_{j=1}^{i-1} \frac{1}{d_{ij}^2} \tag{4.1}$$

where R is a repertoire with $|R|$ distinct sounds, and d_{ij} is the perceptual distance between sound i and sound j. The perceptual distance between vowels is determined by the position of peaks (resonances) in the vowel's frequency spectrum. The frequency of the first and the second peak can be used as coordinates in a two-dimensional space. The weighted Euclidean distance between two such points turns out to be a good measure of perceptual distance between vowels. E is a measure for the quality of the system, where lower values correspond to a better distinguishable

repertoire. The E stands for "energy", in analogy with the potential energy that is minimised in various models in physics.

Liljencrants & Lindblom (1972) performed computer simulations using a simple hill-climbing heuristic, where at each step a random change to the repertoire is considered, and adopted only if it has a lower energy than the current state. Their results compare favourably to observed data on vowel system distributions. In figure 4.1, I show similar results from a simple model with an abstract acoustic space that is a simple 1×1 square.



(a) 5 signals (b) 9 signals (c) 40 signals

Figure 4.1: Configurations of 5, 9 and 40 signals in an abstract acoustic space, when optimised according to the Liljencrants & Lindblom (1972) criterion (equation 4.1). Shown are the configuration found after 2000 hillclimbing steps, with distortions drawn from a Gaussian distribution around each point ($\sigma_x = \sigma_y = 0.05$), starting from initial configurations where each signal has (uniformly) random coordinates.

Liljencrants & Lindblom's results were important because they showed that the configuration of sound systems in natural languages is not arbitrary, and can be understood as the result of more fundamental principles. However, a number of questions remain. First of all, what in the real world exactly is the optimisation criterion meant to be modelling? The Lindblom & Liljencrants model is often described as "maximising the distances between vowels" or "minimising the probability of confusion". It is important to realise, however, that the optimisation criterion in equation (4.1) is neither. Minimising $E = \frac{1}{2}\sum_{i,j\neq i}^{N} 1/d_{ij}^2$ by changing the configuration of a set of vowels in a restricted acoustic space, is not necessarily the same a maximising the average distance $\overline{d} = \frac{1}{N}\sum_{i,j\neq i}^{N} d_{ij}$ (or squared distance), nor is it the same as minimising the average confusion probability $\overline{C} = \frac{1}{N}\sum_{i,j\neq i}^{N} P(j\ perceived|i\ uttered)$. At intermediate distances, these three criteria behave very similarly. The crucial difference is at distance 0, when Liljencrants & Lindblom's E goes to infinity, and at large distances, when both the E and $\overline{C}$ measures, but not $\overline{d}$, approach 0. In section 4.3.2 I will argue that in some cases Liljencraft and Lindblom's E in fact

behaves unrealistically and that minimising the average confusion probability (or equivalently, maximising the *distinctiveness* $D = 1 - \overline{C}$) is a better criterion.

Second, we should ask which mechanism in the real world is responsible for the optimisation? Lindblom himself has referred to both natural selection and self-organisation. As is discussed in chapter 2, the frequency dependence of language evolution means that natural selection on the level of the individual cannot be equated with optimisation on the level of the population. Before we can invoke natural selection, we need to do at least a game-theoretic analysis to show that each new configuration of signals in the acoustic space can *invade* in a population where it is extremely rare. Models of this type will be discussed in the next section.

For self-organisation, the mechanism for optimisation has been worked out more precisely. De Boer (2000; 2001) has studied a simulation model of a population of individuals that each strive to imitate the vowels of others, and be imitated successfully by others. The agents in the model have simplified but realistic mechanisms for recognition and articulation of vowels. They maintain a repertoire of vowel prototypes, and modify the repertoire depending on their success in imitating the vowels of others, as well as the success in being imitated. De Boer showed that in this process of self-organisation – where all agents learn from each other – similar configurations of vowels emerge as in the Liljencrants & Lindblom (1972) model, and as found empirically in the languages of the world. De Boer's model does not explain, however, where the specific learning procedures come from.

Finally, the important question remains of how to extend these models to more complex signals? The models of Liljencrants, Lindblom and De Boer only deal with vowels and, hence, only with the *discrete* aspect of human phonology. They have little to say about the evolution of superficial and productive combination. Lindblom *et al.* (1984), and similarily de Boer (1999, chapter 7), have studied models where signals are trajectories, going from a point in a consonant space, to a point in a vowel space. But in these models the issue still really is the emergence of categories, because the sequencing of sounds is taken as given and there is no interaction between the dynamics in consonant space and those in vowel space.

4.2.2 *Natural selection for combinatorial phonology*

Nowak & Krakauer (1999) apply notions from information theory and evolutionary game theory to the evolution of language. They derive an expression for the "fitness of a language". Imagine a population of individuals, a set of possible signals and a set of possible meanings to communicate about. Speakers choose a meaning to express (the intention), and choose a signal for it with a certain probability. Hearers receive the signal, possibly distorted due to a certain degree of noise.

Hearers subsequently decode the (distorted) signal and arrive at an interpretation. Using similar notation as in section 3.5 of the previous chapter, the *payoff* of the interaction between a speaker and a hearer is described with the equation:

$$w = \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{S}_{mi} \mathbf{U}_{ij} \mathbf{R}_{jm}), \tag{4.2}$$

where w is the expected payoff from an interaction between a speaker and a hearer, $\mathbf{S}_{mi}$ gives the probability that the speaker will use signal i to express meaning m, $\mathbf{U}_{ij}$ gives the probability that i is perceived as j, and $\mathbf{R}_{jm}$ gives the probability that the hearer interprets j as m. This equation is identical to equation 3.1 when all possible meanings are equally valuable and equally frequent (that is, the reward matrix $\mathbf{V}$ is the identity matrix).

Nowak & Krakauer define a language as the combination of a production and interpretation matrix, i.e. $L = \{\mathbf{S}, \mathbf{R}\}$. The fitness of a language L in a situation where one needs to communicate with users of a language L' (where $L' = \{\mathbf{S}', \mathbf{R}'\}$), is now given by (assuming speaking and hearing are equally important):

$$F(L, L') = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathbf{S}_{mi} \mathbf{U}_{ij} \mathbf{R'}_{jm} + \mathbf{S'}_{mi} \mathbf{U}_{ij} \mathbf{R}_{jm}\right). \tag{4.3}$$

Nowak et al. observe that when communication is noisy and when a unique signal is used for every meaning, the fitness is limited by an "error limit": only a limited number of sounds can be used – and thus a limited number of meanings be expressed – because by using more sounds the successful recognition of the current signals would be impeded. Nowak et al. further show that in such noisy conditions, fitness is higher when (meaningless) sounds are combined into longer words. When the environment is combinatorial (i.e. objects and actions occur in many combinations) the fitness is highest when meaningful words are combined into longer sentences (compositionality). These results are essentially particular instantiations of Shannon's more general results on "noisy coding" (Shannon, 1948), as is explored in a later paper by the same group (Plotkin & Nowak, 2000).

More interesting is the question how natural selection could favour a linguistic innovation that introduces combinatorial phonology, in a population where that innovation is still very rare (the "invasibility requirement", discussed in chapter 2). Nowak & Krakauer (1999) do not address that specific problem mathematically. They do, however, perform a mathematical, game-theoretic analysis of the evolution of "compositionality", and point out that this analysis can be easily

adapted to the case of combinatorial phonology. It is worth spelling out the analysis for combinatorial phonology, because it reveals some strong assumptions.

In the analysis of compositionality, all mixed strategies are considered where both holistic and compositional signals are used. Nowak & Krakauer show that strategies that use more compositionality can invade strategies that use less. This means that under natural selection, languages should evolve compositionality. When applied to combinatorial phonology[3], the analysis starts with **S**- and **R**-matrices of the following form:

$$\mathbf{S} = \begin{array}{c|cccccccc} & a & b & c & d & AA & AB & BA & BB \\ \hline m_1 & 1-x & 0 & 0 & 0 & x & 0 & 0 & 0 \\ m_2 & 0 & 1-x & 0 & 0 & 0 & x & 0 & 0 \\ m_3 & 0 & 0 & 1-x & 0 & 0 & 0 & x & 0 \\ m_4 & 0 & 0 & 0 & 1-x & 0 & 0 & 0 & x \end{array}, \quad \mathbf{R} = \begin{array}{c|cccc} & m_1 & m_2 & m_3 & m_4 \\ \hline a & 1 & 0 & 0 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 0 \\ d & 0 & 0 & 0 & 1 \\ AA & 1 & 0 & 0 & 0 \\ AB & 0 & 1 & 0 & 0 \\ BA & 0 & 0 & 1 & 0 \\ BB & 0 & 0 & 0 & 1 \end{array},$$

where x is a single variable that describes how often the holistic strategy is used (with signals a, b, c, d) vs. how often the combinatorial strategy is used (with words built-up out of the phonemes A and B). Nowak & Krakauer further assume that the confusion between holistic signals (u_h) is larger than the confusion between words (u_c), and that there is no confusion between the two types of strategies. Hence, if we write out the resulting confusion matrix **U**, it looks like this:

$$\mathbf{U} = \begin{array}{c|cccccccc} & a & b & c & d & AA & AB & BA & BB \\ \hline a & u_h & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 \\ b & \cdot & u_h & \cdot & \cdot & 0 & 0 & 0 & 0 \\ c & \cdot & \cdot & u_h & \cdot & 0 & 0 & 0 & 0 \\ d & \cdot & \cdot & \cdot & u_h & 0 & 0 & 0 & 0 \\ AA & 0 & 0 & 0 & 0 & u_c & \cdot & \cdot & \cdot \\ AB & 0 & 0 & 0 & 0 & \cdot & u_c & \cdot & \cdot \\ BA & 0 & 0 & 0 & 0 & \cdot & \cdot & u_c & \cdot \\ BB & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & u_c \end{array},$$

where the entries marked with a '$\cdot$' can be ignored because they will be multiplied with 0.

[3]Note that, even though we are only interested in phonology here, "meanings" do have a role because they determine how many signals have to be kept distinct.

With these assumptions, it is straightforward to describe the fitness of speakers of languages L and L' when communicating with their own type or with the other, using equation (4.3) (many terms simplify because of the many zero's in **S**, **R** and **U**). If L has a tendency x to use the combinatorial strategy, and L' a tendency x', these fitnesses are:

$$F(L,L) = M\left((1-x)\,u_h + xu_c\right) \tag{4.4}$$

$$F(L,L') = \frac{1}{2}M\left(\left((1-x)\,u_h + xu_c\right) + \left(\left(1-x'\right)u_h + x'u_c\right)\right) \tag{4.5}$$

$$F(L',L') = M\left(\left(1-x'\right)u_h + x'u_c\right). \tag{4.6}$$

From these equations, it follows immediately that a more combinatorial language can always invade a population with a less combinatorial language. Note the following inequalities:

$$F(L',L') > F(L,L') > F(L,L) \;\; if \;\; (x' > x) \wedge (u_c < u_h). \tag{4.7}$$

This satisfies the criterion of invasibility discussed in chapter 2. If L' is very infrequent, then all speakers of language L (the "residents") will have a fitness of approximately $F(L,L)$ and the rare speakers of language L' (the "mutants") will have fitness of approximately $F(L,L')$, because for both residents and mutants the vast majority of interactions will be with speakers of language L. Once the frequency of mutants starts to rise, the residents will gain in fitness, that is, move toward a fitness $F(L,L')$. However, the mutants will gain even more by interacting more and more with other mutants, that is, move toward $F(L',L')$.

Although this model is a useful formalisation of the problem and gives some important insights, as an explanation for the evolution of combinatorial phonology (and compositionality) it is unconvincing. The problem is that the model only considers the advantages of combinatorial strategies, and ignores two obvious disadvantages: (1) by having a "mixed strategy" individuals have essentially two languages in parallel, which one should expect to be costly because of memory and learning demands and additional confusion[4]. Nowak & Krakauer simply assume that the second system is in place, and that the hearer interprets all signals correctly, even if x is close to 0, and the number of learning experiences is therefore extremely small; (2) combinatorial

[4]An interesting question is what exactly the costs of speaking are. Modern humans acquire and use their native language without much effort, and with neglible energy use (Fitch & Hauser, 2002). However, that does not necessarily generalise to earlier stages of language, and to the different variants that we consider in evolutionary models. Moreover, there always are, as I argued in chapter 2, biophysical constraints on the available strategies. What counts in evolutionary modelling is not an absolute measure of costs, but rather the relative advantages and disadvantages of the competing strategies. The best interpretation of "costs" here, is therefore probably "missed opportunity costs".

signals that consist of two or more sounds take longer to utter and are thus more costly[5]. A fairer comparison would be between holistic signals of a certain duration (where continuation of the same sound decreases the effect of noise) and combinatorial signals of the same duration (where the digital coding decreases the effect of noise). This is the approach I take in the model of this chapter, but like Nowak & Krakauer, I will look at invasibility in addition to optimisation.

4.2.3 *Crystallisation in the perception–imitation cycle*

A completely different approach to combinatorial phonology is based on "categorical perception". Categorical perception (Liberman *et al.*, 1957; Harnad, 1987) is the phenomenon that categorisation influences the perception of stimuli in such a way that differences between categories are perceived as larger and differences within categories as smaller than they really are (according to an "objective", cross-linguistic similarity metric). For instance, infants of just a few months old already perceive phonemes as closer to the closest prototype phoneme from their native language than it is according to an "objective" (cross-linguistic) acoustical metric (Kuhl *et al.*, 1992). Apparently, the frequency and position of acoustic stimuli gives rise to particular phoneme prototypes, and the prototypes in turn "warp" the perception.

Oudeyer (2001, 2002) observes that signals survive from generation to generation because they are perceived and imitated. Categorical perception will therefore *shape* a signal repertoire such that it conforms more and more to the prototype phonemes. Thus, emitted signals shape perception, and distorted perception shapes the repertoire of signals in the cycle from emission to perception to emission (the perception–imitation cycle; see also Westermann, 2001, for a model of sensori-motor integration and its relevance for imitation and categorical perception).

Oudeyer (2001) presents a model to study this phenomenon. In this model, signals are modelled as points in an acoustic space. The model consists of two coupled neural maps, one for perception and one for articulation. The perceptual map is of a type known to be able to model categorial perception: its categorisation behaviour changes in response to the input data. It is sketched in figure 4.2. In addition, the associations between perceptual stimuli and articulatory commands are learned. Through this coupling between perceptual and articulatory maps, a positive feedback loop emerges where slight non-uniformities in the input data lead to clusters in the perceptual map, as well as weak clusters in the articulatory map, and hence to slightly stronger

[5]It is, of course, slightly awkward to criticise a model that Nowak & Krakauer (1999) never actually worked out. The point here is that if one takes all the assumptions that they do spell out in the paper, and work out the model as they suggest, the result is unsatisfactory. A better model of the evolution of combinatorial phonology must start with different assumptions.

non-uniformities in the distribution of acoustic signals. Oudeyer calls the collapse of signals in a small number of clusters "crystallisation".

Oudeyer (2002) generalised these results to a model with (quasi-) continuous trajectories, where a production module triggers a sequence of targets in the articulatory map, which yield a continuous trajectory. This trajectory is then discretised at a very fine sampling rate, and each point is presented to the neural map as before. Also in this version of the model, well-defined clusters form in the perceptual and articulatory maps. The signals can thus be analysed as consisting of sequences of phonemes.

Oudeyer's model is fascinating, because it gives a completely non-adaptive mechanism for the emergence of combinatorial phonology. However, the model does make a number of important assumptions, such as the pressure for vocal imitation (a skill that is in fact very rare among primates; Fitch, 2000) and the parameters of the neural maps. The validity of those assumed traits would be much strengthened if one could show that these traits would be selected for in evolution. It therefore remains an important question whether recombination increases the functionality of the language, and thus the fitness of the individual that uses it. If not, one would expect selection to work against it.

In particular, in Oudeyer's first model (2001), where signals are instantaneous, a large repertoire of signals is collapsed into a small number of clusters. A functional pressure to maintain the number of distinct signals would thus have to either reverse the crystallisation, or combine signals from different clusters. In his second model (Oudeyer, 2002), signals are continuous trajectories and potentially a much larger distinct repertoire can emerge. However, the functionality of the repertoire is not monitored, and plays no role in the dynamics. It might or might not be sufficient. The number of "phonemes" (the discrete aspect) that forms is a consequence of the parameters and initial configuration, and in a sense accidental. The reuse (the superficial combination aspect) in the model is built-in in the production-procedure.

The assumption that signals already consist of sequences of articulatory targets is justified with considerations from articulatory phonetics, as I discussed in chapter 3, section 3: constraints from articulatory motor control, it is argued, impose combinatorial structure on any large repertoire of distinct sounds. Even if one fully accepts this argument, the need for a large and distinctive repertoire is a functional pressure. In Oudeyer's model, however, there is no interaction between the number of phonemes that is created, and the degree of reuse (the number of phonemes per signal) that emerges. This issue, which seems the core issue in understanding the evolutionary

origins of combinatorial phonology, is not modelled by Oudeyer. In my model, in contrast, I ensure that the functionality increases rather than decreases.

4.2.4 *Other models*

All other models of the evolution of combinatorial phonology that I am aware of, also assume the sequencing of phonetic atoms into longer strings as given. They concentrate rather on the structure of the emerged systems (Lindblom *et al.*, 1984; de Boer, 2001; Redford *et al.*, 2001) or on how conventions on specific combinatorial signal systems can become established in a population through cultural transmission (Steels & Oudeyer, 2000). Theories on the evolution of speech developed by linguists and biologists focus on possible pre-adaptations for speech. MacNeilage & Davis (2000) propose oscillatory movements of the jaw such as used in breathing and chewing as precursors for syllable structure. Fitch (2000) sees sexual selection as a mechanism to explain the shape of the human vocal tract. Studdert-Kennedy (2002) explains the origin of recombination and duality of patterning as the result of vocal imitation. Finally, connectionist models of phoneme discovery (e.g. Kohonen, 1988; Waibel, Hanazawa, Hinton, Shikano & Lang, 1989; Guenther & Gjaja, 1996) learn from a samples from a language that already shows combinatorial phonology.

These models and theories are interesting, and, importantly, bridge the gap with empirical evidence on how combinatorial phonology is implemented in the languages of the world. However, they are of less relevance here, because they do not address the origins of the fundamental, qualitative properties of discrete and combinatorial coding. That is, they leave open the question as to under what circumstances a system of holistically coded signals with finite duration would change into a combinatorial system of signals.

4.3 Model Design

I will now present the design of a new model, that shares features with all three existing approaches. Like Liljencrants & Lindblom (1972) and other models, it makes use of the concept of "acoustic space", a measure for perceptual distinctiveness and a hill-climbing heuristic. Like the Nowak & Krakauer (1999) model, the measure for distinctiveness is based on confusion probabilities, and my study includes a game-theoretic invasibility analysis. Finally, like Oudeyer (2002), I model signals not just as points, but as trajectories through acoustic space.

In the model, I do not assume combinatorial structure, but rather study the gradual emergence of superficially combinatorial phonology from initially holistic signals. I do take into account the temporal structure of both holistic and phonemically coded signals. I view signals as continuous

movements ("gestures", "trajectories") through an abstract acoustic space. I assume that signals can be confused, and that the probability of confusion is higher if signals are more similar, i.e. closer to each other in the acoustic space according to some distance metric. I further assume a functional pressure that maximises distinctiveness.

4.3.1 The acoustic space

The model of this chapter will deal with repertoires of signals, their configuration and the similarities between signals. This requires conceptualising signals as points or movements through a space. How could we define such an "acoustic space"? An appropriate definition of acoustic space will, as much as possible, reflect the articulatory constraints as well as perceptual similarities, such that signals that cannot be produced fall outside the space, and that points in the space that are close sound similar and are more easily confused.

For vowels, a simple but very useful acoustic space can be constructed by just looking at the peaks in the frequency spectrum. These peaks (called "formants") correspond to the resonance frequencies in the vocal tract, and are also perceptually very salient. Artificially produced vowels with the correct peaks but otherwise quite different frequency spectra, are recognisable by humans. From experiments where subjects are asked to approximate vowels sounds by manipulating just two formant frequencies, it is clear that a good representation of vowels can be given in just two dimensions, with the first formant as the first dimension, and the *effective second formant* as the second dimension (Carlson *et al.*, 1970). There are a number of simple formulas (e.g. Mantakas, Schwartz & Escudier, 1986, discussed in de Boer 1999) for calculating the effective second formant, F_2', given the second, third and fourth formant frequencies F_2, F_3 and F_4 (measured in "Barks", a scale based on psychophysical experiments, Traunmüller, 1990).

Perceptual distance between two vowels a and b in the space of first and effective second formant is typically defined as the (weighted) Euclidean distance:

$$d(a,b) = \sqrt{\left(F_1^a - F_1^b\right)^2 + \lambda^2 \left(F_2^a - F_2^b\right)^2}, \tag{4.8}$$

where λ is a weight that balances the importance of the effective second formant relative to the first formant, which is experimentally estimated at $\lambda = 0.3$ (see de Boer, 1999 and references therein).

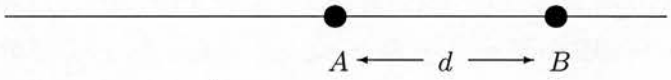
A related approach for defining an acoustic space works with *cepstral coefficients* (Bogert, Healy & Tukey, 1963). These coefficients (obtained by taking the inverse Fourier Transform of the log of the frequency spectrum) describe the general shape of the frequency spectrum. The first

coefficient is a measure of the total energy of the signal, and the subsequent coefficients give more and more detailed information about the signal. The cepstral coefficients thus define a sequence of features of the signal of decreasing importance. Vowels and diphthongs can be accurately represented with the first two coefficients; for consonants we need five or six.

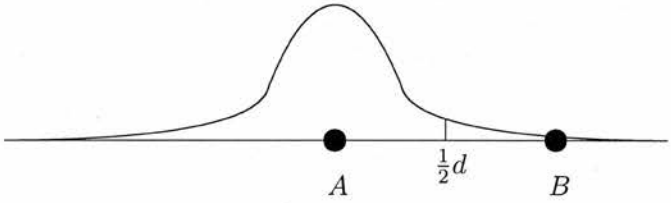
Although formants work well for humans, and in particular for European languages, pitch is a more salient variable in the vocalizations and perception of other animals (although it is now believed that several mammal species are able to perceive formants as well, e.g. Reby, McComb, Cargnelutti, Darwin, Fitch & Clutton-Brock, 2005). Of course, it is difficult to tell what the appropriate acoustic space is for modelling articulation and perception of early hominids that feature in scenarios of the evolution of language (e.g Lieberman, 1984; Jackendoff, 2002). However, the considerations that will be presented below remain the same, independent of the exact nature of the underlying perceptual dimensions.

4.3.2 *Confusion probabilities*

Once we have constructed an acoustic space that captures the notion of perceptual similarity, we can ask how distance in that space relates to the probability of confusion? Answering that question requires us to make assumptions about the causes of confusion and the nature of categorisation. We can get a general idea, by first looking at the simple example of a 1-dimensional acoustic space, with just 2 prototype signals A and B (modelled as points in that space), and a distance d between them:



Now assume that a received signal X, lying somewhere on the continuum, will be perceived as A or B depending on which is closest (*nearest neighbour classification*). Finally, assume a degree of noise on the emitted signals, such that when a signal, say A, is uttered, the received signal X is any signal drawn from a Gaussian distribution around A:



Now we can calculate the probability that an emitted signal A is perceived as B:

$$\begin{aligned} P(B \textit{ perceived}|A \textit{ uttered}) &= \int_{x=\frac{1}{2}d}^{\infty} \mathcal{N}(\mu = 0, \sigma = \delta)\,\mathrm{d}x \\ &= \int_{x=\frac{1}{2}d}^{\infty} \frac{1}{\sqrt{2\pi}\delta} e^{\frac{-x^2}{2\delta^2}}\,\mathrm{d}x, \end{aligned} \tag{4.9}$$

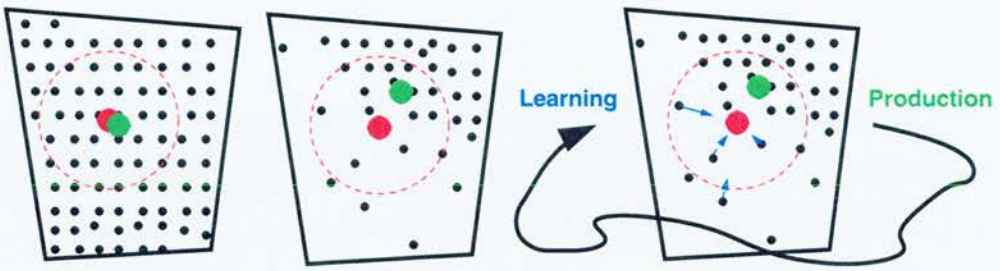
where δ is the standard deviation[6] of the Gaussian, and hence the *characteristic distance* of the noise function. This integral, which describes the area under the Gaussian curve right of the point $\frac{1}{2}d$, can be solved numerically, as in figure 4.3 (the integral of the Gaussian is called the "error function": $\mathrm{erf}(x)$).

This function has a number of important features. First of all, if the two signals A and B are identical (i.e. $d = 0$), the confusion probability is not 100%, as the naive first intuition might be, but 50%. Second, with increasing d, the confusion probability first rapidly decreases and then slowly approaches 0 (see figure 4.3). Of course, the confusion probability as a function of distance can have many different shapes depending on the exact type of noise and the exact type of categorisation. Ultimately, this is an empirical question. It seems, however, that the function will always have these general characteristics at $d = 0$ and in the limit of $d \to \infty$.

In contrast, the previously discussed E measure (Liljencrants & Lindblom, 1972), and average distance measure $\overline{d}$, do not have both these properties. For the purposes of this chapter they are therefore not appropriate criteria for optimisation. Figure 4.1(c) serves as an illustration: here many signals are crammed into a small space. The configuration that maximises E will, regardless of the noise level, always keep all signals distinct. The configuration that maximises $\overline{d}$ will, again regardless of the noise level, always collapse all signals in 4 clusters in the four corners of the space. At intermediate noise levels, both configurations are in fact suboptimal. A given noise level defines a "channel capacity" for the acoustic space: cramming more or fewer signals in the space will in fact impede the amount of information that can be encoded (Shannon, 1948).

If the acoustic space has more than 1 dimension, and if there are more than 2 signals, calculations like in equation (4.11), quickly get extremely complex, and confusion probabilities are no longer uniquely dependent on distance. We can, however, assume that the confusion probabilities are generally proportional to a function of distance with a shape as in figure 4.3. Hence, let $f(d)$

[6] To avoid confusion, I will use the symbol δ for the standard deviation (characteristic distance) of the acoustic noise function, ρ for the standard deviation (the hill-climbing rate) of the hillclimbing heuristic that will be introduced later, and σ for standard deviations in general.



(a) Each neuron in the perceptual map responds maximally to sounds at a specific point in the acoustic space (the "location" of that neuron, drawn as black points in the graph), and with decreasing strength to sounds that are further away. The response (drawn as a green circle) of the perceptual map to a given signal (the red circle) is calculated as a weighted average of the locations of neurons. If neurons are distributed uniformly over the acoustic space (left panel) the response is accurate; if they are distributed non-uniformly (right panel) perception is warped.

(b) In learning, the "location" of neurons is shifted towards the given signal (blue arrows). Over time, the perceptual map will therefore reflect the distribution from which the signals are drawn. The response of the map will now be warped towards the most frequently observed signals, which constitutes a form of categorial perception. In the model of Oudeyer (2001), the distribution of perceived signals is again dependent, through the coupling of the perceptual and articulatory maps, on the existing categorical perception. A slight effect of categorical perception will therefore be reinforced, until very strong clusters of "phonemes" emerge.

Figure 4.2: Perceptual Warping, Categorical Perception and the emergence of combinatorial phonology.

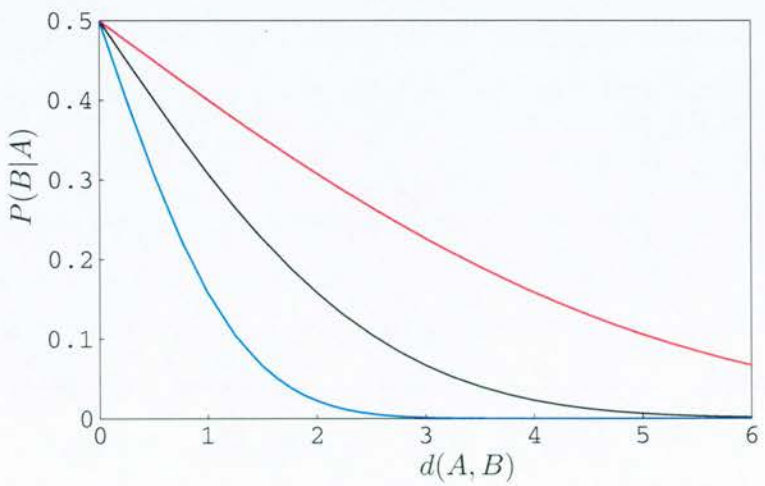


Figure 4.3: The probability of confusion as a function of distance (numerical solutions to equation (4.9) for $\delta = 0.5$ (bottom curve), $\delta = 1$ (middle curve), and $\delta = 2$ (top curve)).

be a function of distance d of that shape, parametrised by noise level δ :

$$f(d) = \int_{x=\frac{1}{2}d}^{\infty} \frac{1}{\sqrt{2\pi}\delta} e^{\frac{-x^2}{2\delta^2}} \, \mathrm{d}x. \tag{4.10}$$

I will call the result $f(d)$ of a specific d the "f-score" of d. As a first approximation, we can assume that confusion probabilities are proportional to these f-scores: $P(B \textit{ perceived}|A \textit{ uttered}) \propto f(d(A, B))$. But we also know that the probabilities of confusing a signal with any of the other signals in a repertoire (including the signal itself) must add up to 1: $\sum_{X \in R} P(X \textit{ perceived}|A \textit{ uttered}) = 1$. Hence, we can estimate the probability of confusing signal A with signal B as:

$$P(B \textit{ perceived}|A \textit{ uttered}) = \frac{f(d(A, B))}{\sum_{X \in R} f(d(A, X))}. \tag{4.11}$$

From this, it is now straightforward to define a measure for the *distinctiveness* $D(R)$ of a repertoire R with T signals. Let D be the average probability that a signal R_t from R is correctly identified (assuming all signals are used equally often):

$$D(R) = \frac{1}{T} \sum_{t=1}^{T} \frac{f(d(R_t, R_t))}{\sum_{t'=1}^{T} f(d(R_t, R_{t'}))}. \tag{4.12}$$

This equation relates the distances between signals in acoustic space to the probability of confusing a random signal with another one, and thus gives a quality measure for repertoires of signals. The measure plays an important role in the model I present here, even for more complex signals that I will consider below. Ultimately, such a measure should of course be based on empirical findings. However, the results I will present in this chapter do not depend on the exact properties of this measure. I will come back to this issue later in this thesis.

4.3.3 Trajectory representation

We have a qualitative understanding of how to define the acoustic space with points representing signals, and of how to estimate the confusion probabilities and distinctiveness as functions of the distances between those points. We can now try to extend the model to deal with signals that have a temporal dimension. It would be desirable if the same apparatus can still be used. I therefore define temporal signals as *trajectories*: movements through the acoustic space. In my approach, a trajectory is a connected sequence of points (each of which could correspond to, for instance, the cepstral coefficients of the frequency spectrum of a small interval in the waveform).

To illustrate the feasibility of deriving trajectory representations from acoustic data, I show in figure 4.4(a) a number of trajectories through vowel space that are based on actual recordings.

The graph shows the trajectories from a number of recorded vowels, which correspond to more-or-less stationary trajectories in the space, and from recordings of a number of diphthongs, which correspond to movements from one vowel's region to another. Figure 4.4(b) and (c) show trajectories through the space defined by three of the first 6 cepstral coefficients. In this space we can, to a certain extent, represent consonants as well. Overlaid in both graphs are the resulting trajectories of two recordings, illustrating that the construction is repeatable, albeit with considerable variation.

In the model of this chapter I will not worry about the problems of constructing acoustic spaces and drawing trajectories through them. Instead, I will take as the starting point a set of trajectories through an abstract acoustic space. The model is based on piece-wise linear trajectories in bounded 2-D or 3-D continuous spaces of size 1×1 or $1 \times 1 \times 1$. Trajectories are sequences of points with fixed length (parameter P), that always stay within the bounds of the acoustic space. In the standard model, each point has a fixed distance (parameter S) to the immediately preceding and following points in the sequence. I will also consider a variant, where this distance is not fixed but either completely unconstrained, or anything between 0 and a given maximum value. That is, if t_x and t_{x+1} are neighbouring points, t_{x+1} can lay anywhere within a circle of the given radius around t_x (in the graphs this is indicated as "segment size unconstrained" or, e.g., "$S \leq 0.1$").

Signals in the real world are continuous trajectories, but in the model I need to discretise the trajectories. To ensure that I do not impose the combinatorial structure we are interested in, I discretise at a much finer scale than the phonemic patterns that will emerge. Hence, the points on a trajectory are not meant to model atomic units in a complex utterance. They implement a discretisation of a continuous trajectory that can represent either a holistically coded or phonemically coded signal.

4.3.4 *Measuring distances and optimising distinctiveness between trajectories*

How do we extend the distance and distinctiveness measures for points to trajectories? Perhaps the simplest strategy would be to look at a repertoire of trajectories one time-slice at a time, and simply optimise – as before – the distinctiveness between the points. This is similar to the approach in Lindblom *et al.* (1984). With such an approach, however, the temporal dimension could just as well have been left out, and the model has little to say about the emergence of superficial combination. Combinatorial phonology emerges – if at all – as a trivial consequence of (i) the formation of categories (phonemes), and (ii) sequencing imposed by the trajectory representation. Whether or not signals are repetitions of the same phoneme, or combinations of different

phonemes, depends on the initial configuration and the possible constraints on the shape of trajectories. This is illustrated in figure 4.5.

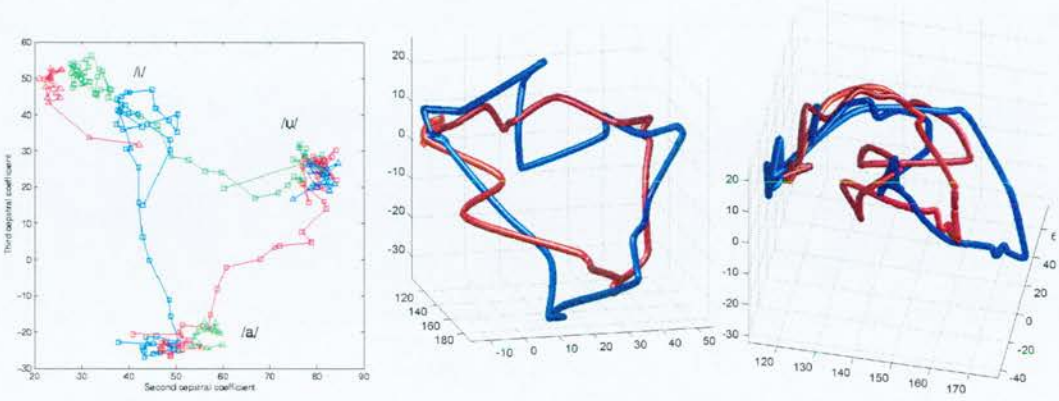
Much more interesting is when we measure the distance between complete trajectories and optimise their distinctiveness. In such an approach, there is a role for combinatorial phonology: the confusion probability between two largely overlapping trajectories might be very low, as long as they are sufficiently distinct along one stretch of their length. As a provisional measure, I define the distance between two trajectories t^i and t^j, as the *average* distance between the corresponding points on the trajectories:

$$d(t^i, t^j) = \frac{1}{P} \sum_{p=1}^{P} d(t_p^i, t_p^j), \tag{4.13}$$

where t_p^i is the p-th point on the i-th trajectory in a repertoire, and $d(a, b)$ gives the distance between two points a and b. This distance measure then provides the input to the same of distance-to-confusion function that I derived for points (equation 4.12).

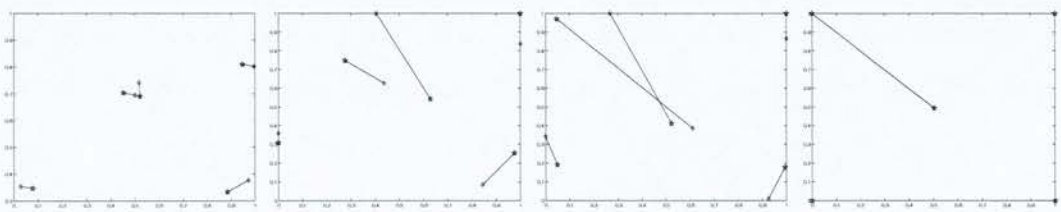
One may argue that the distance metric in equation (4.13) is too simplistic, and does not do justice to the fact that slight differences in timing of two signals will not affect their perceptual similarity much. One should thus expect a high probability of confusion between two such signals, even though according to equation (4.13) they are very far apart. An alternative distance measures, that does take into account such timing effects, is "dynamic time warping" (DTW). DTW is an efficient method that before the advent of statistical models has been used with reasonable success in computer speech recognition (e.g. Sakoe & Chiba, 1978). The distance between two trajectories t and r is then defined as the sum of the distances between all corresponding points in *the best possible alignment* of the two trajectories. In finding the best possible alignment, one point from t can be mapped on several neighbouring points in r and vice versa. In this way trajectories that resemble each other in shape, but that do not align perfectly are still considered close. DTW models the way humans perceive signals, and in the final part of this chapter I will look at a simulation that uses it.

However, even with such an improved distance measure, it is not clear how accurate the estimate of confusion probabilities is. The approach I have adopted here uses the *pairwise distances* of all signals as an intermediate step in going from the *configuration* of trajectories to their *confusion probabilities*. For trajectories, it is far from trivial to derive the exact shape of the distance-to-confusion, even if the noise and categorisation mechanisms were completely known. More work is needed – both empirical and theoretical – to study whether this approach yields realistic results. For the purposes of this chapter, however, I will take a pragmatic approach. It seems,



(a) 3 vowels and 3 diphthongs (b) two recordings of "een" (c) two recordings of "negen"

Figure 4.4: Trajectory representations derived from recorded acoustic data. Each point on a trajectory is given by the cepstral coefficients of the frequency spectrum of a short time interval of the signal. In (a) the first two coefficients are used; in (b) and (c) coefficients 1, 3 and 5 are used. The blue and red curves in (b) and (c) are based on two different recordings each. The graphs illustrate that it is possible to construct an acoustic space for trajectories with a meaningful interpretation. However, it is also clear that there is much variation between different recordings, suggesting that much more work is needed for the trajectory representation to be useful in any practical applications (see Goldenthal, 1994, and subsequent work on using trajectory representations in speech recognition). (Graphs created by Bart de Boer).



(a) t=0 (b) t=100 (c) t=200 (d) t=1300

Figure 4.5: Optimising distinctiveness in each time-slice independently. Shown are the initial configurations of 5 trajectories (each consisting of just 2 points), two intermediate stages, and the stable equilibrium. End-points are marked with a star symbol. The results can be interpreted as "combinatorial phonology" (2 trajectories share begin- end endpoints), but this is a trivial result from the fact that the begin-points and the end-points move independently to their equilibrium configuration, although through chance, the begin- and end-points of each particular trajectory can end up at different locations.

as I argued above, that the *general shape* of the distance-to-confusion function is appropriate. I will therefore mostly use the simple distance measure of equation (4.13), and consider several alternative distance-to-confusion functions to ensure that the results do not crucially depend on this assumption.

4.3.5 The hill-climbing heuristic

Now that I have defined a distance metric, it is straightforward to use a hill-climbing heuristic such as Liljencrants & Lindblom (1972) and apply it to much more complex signals. Hill-climbing is an iterative procedure, where at each step a random change to the repertoire is considered, and if it improves the distinctiveness it is applied. Then another random change is considered and the same procedure applies over and over again. In pseudo-code, the procedure looks as follows:

```
% R is a repertoire of signals
% S is the segment length parameter
% ρ is the hill-climbing rate parameter
% δ is the acoustic noise parameter (characteristic distance)
for i = 1 to I
    R' = CONSTRAIN( R + DISTURBANCE( ρ ), S);
    if D(R',δ) > D(R,δ) then R = R';
end for
```

Here, DISTURBANCE applies random noise ($\mathcal{N}(\mu = 0, \sigma = \rho)$), to all of the coordinates of a (uniformly) random point on a random trajectory. D is the distinctiveness function given in equation (4.12). CONSTRAIN is a function that enforces that all points on the trajectories fall within the boundaries of the acoustic space, and that all segments have maximum length S. Hence, after a random point t_x is moved to a new random position, the CONSTRAIN function first moves it back, if necessary, within the boundaries of the acoustic space; and then, it moves the two points on both sides of the moved point, t_{x+1} and t_{x-1}, if necessary, such that the distance to t_x equals S. The direction from t_x to t_{x+1} or t_{x-1} remains the same, unless the point would cross the boundary of the space, in which case it is placed at a random point within the boundary at distance 1 from t_x. The same procedure is applied recursively to the neighbours of t_{x+1} and t_{x-1} until the ends of the trajectory are reached.

Hill-climbing is just an optimisation *heuristic*; there is no guarantee that it will find the optimal configuration for the given criterion. Especially when the repertoire considered is relatively complex, the system is likely to move toward a local optimum. Although better optimisation heuristics exist, this problem is in general unavoidable for systems with so many variables. Also

in Nature, the optimisation of sound systems may not have escaped the problem of local optima; the real optimum is therefore not necessarily interesting for describing the patterns in human speech. Instead, I will concentrate on general properties of the local optima I find, and on the gradual route towards them.

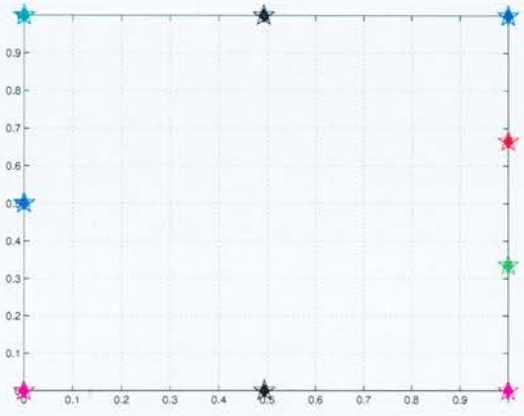
4.4 Results

I have implemented versions of the basic model as outlined above in MatLab. I have run many simulations with a large number of parameter combinations and a number of variations of the basic model. These variations included two alternative noise-to-confusion functions ($f(d) = \frac{1}{1+e^{(\frac{1}{\delta}d)}}$ and $f(d) = \frac{1}{1+e^{(\frac{1}{\delta}d^2)}}$) and an alternative CONSTRAIN function ($S \leq S^*$ and $S = S^*$). Because no real differences were observed in the results, I will here mostly present results with the standard model. In some cases the captions of figures indicate that one of these alternatives is used. In the following I will first briefly give an overview of the general behaviour of the model in these simulations by means of a representative example, and then give a detailed analysis of why I observe the kind of results that I do. In this section I consider simple optimisation; in the next section I will extend these results to a game-theoretic analysis.

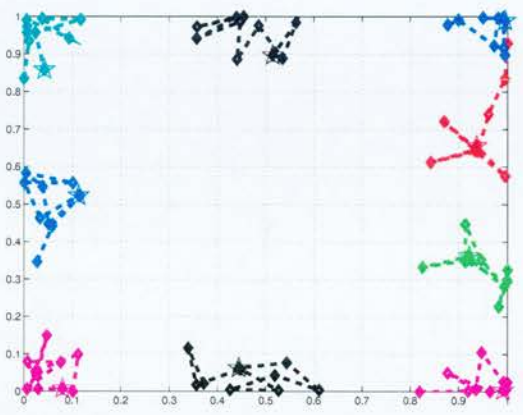
4.4.1 An overview of the results

I will describe the behaviour of the model in many different simulations by using the representative example depicted in figure 4.6. Figure 4.6(a) shows an the equilibrium configuration of 9 point signals in an abstract acoustic space, optimised for distinctiveness at an intermediate noise-level ($\delta = 0.2$). This particular configuration is stable: no further improvements of the distinctiveness of the repertoire can be obtained by making small changes to the location of any of the signals. The locally optimal distinctiveness is $D = 0.66$; that is, with the given noise level, our estimate of the probability of successful communication of a signal is 66%.

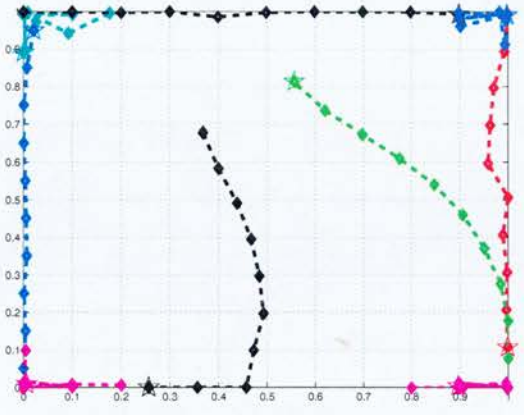
Figure 4.6(b) shows 9 trajectories, consisting of 10 points and hence 9 segments each. Each of these trajectories was created by taking 10 copies of one of the points in figure (a) and connecting them. A small amount of noise was added to each point, and the CONSTRAIN function, as described above, was applied to each trajectory, enforcing a fixed distance ($S = 0.1$) between all neighbouring points on the same trajectory. Due to this perturbation, the distinctiveness of this repertoire of trajectories is somewhat lower, $D = 0.62$, than of the repertoire in (a). (The definitions of distance and distinctiveness are such that a repertoire of stationary trajectories has the same D as a repertoire of points at the same locations; hence, points and stationary trajectories, if all the same length, are equivalent in the basic model).



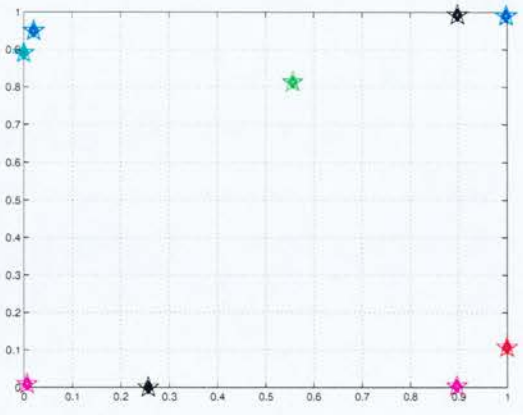
(a) 9 instantaneous signals optimised for distinctiveness. D=0.66. Parameters: $P = 1$.



(b) 9 trajectories at same locations as the signals in (a) with small perturbations. D=0.62. Parameters: $P = 10$, $S = 0.1$.



(c) 9 trajectories optimised for distinctiveness. D=0.69. Parameters: $P = 10$, $S = 0.1$, $\rho = 0.1$, $I = 5000$.



(d) 9 instantaneous sounds at same location as endpoints in (c). D=0.55. Parameters: $P = 1$.

Figure 4.6: In a combinatorial phonology, distinctiveness of signals at each particular time-slice is sacrificed for better distinctiveness of the whole trajectory. Instantaneous signal (or equivalently, stationary trajectories) will be organised in patterns like (a) and not like (d) when optimised for distinctiveness. For non-stationary trajectories, the same pattern, as in (b), is not stable, but will – after optimisation– instead be organised like (c). Each individual time-slice, as illustrated with the end-points in (d) is suboptimal, but the whole temporal repertoire is at a local optimum. Common parameters: $T = 9$, $\delta = 0.2$

What will happen if we now optimise, through hill-climbing, the repertoire of trajectories for distinctiveness? One possibility is that the applied perturbations are nullified as much as possible, such that the system moves back to the configuration of (a). That is not what happens, however. Rather, the system moves to a configuration as in figure 4.6(c). In this configuration, there are 3 trajectories along the left, top and right boundaries; there are 4 trajectories bunched up in each of the four corners; and there are 2 trajectories crossing the acoustic space, one starting near the center and ending near the bottom-left corner, and one starting in the bottom-right corner at ending near the center.

This graph shows a number of important features. First, all trajectories start or end near to where other trajectories start and end. The repertoire therefore can be said to exhibit a *superficially combinatorial phonology*: if we label the corners A, B, C and D, and the central region E, starting top-left and going clockwise, we can describe the repertoire as: $\{A, B, C, D, AB, BC\ CE, ED, DA\}$. That is, we need only 5 category labels (phonemes) to describe a repertoire of 9 signals. In contrast, the repertoire in (b) is most easily described by postulating 9 categories, one for each trajectory[7].

Second, some trajectories are bunched up in as small a region as possible, but other trajectories are stretched out over the full length of the space. Third, the configuration of the repertoire appears somewhat idiosyncratic and is in a local optimum[8]. Fourth, at each time-slice the configuration of the corresponding points is in fact suboptimal. For instance, in figure 4.6(d) just the endpoints of the trajectories in (c) are shown. 8 out of 9 of these points are closer to their nearest neighbour than any of the points in (a). Before I extend these results to simulations with many more trajectories of various lengths, and to acoustic spaces with more dimensions, I will first look at a number of simple cases that explain why the optimised repertoires have these features.

4.4.2 *The optimal configuration depends on the noise level*

To evaluate the role of the noise parameter δ, it is instructive to first look at a simple, 1-dimensional example with signals as points. Consider a situation with 3 signals, 2 of which are fixed at the edges of a 1-dimensional acoustic space. The third signal is at distance x from the leftmost signal, and at distance $1 - x$ from the rightmost signal:

[7] I implicitly assume a model of categorisation here that favours robust and coherent categories. An interesting and important question is how to measure the degree of "combinatoriality", but in this thesis I will rely on an intuitive notion. I'll briefly come back to this issue in chapter 7.

[8] The stability of this configuration has not been rigorously established, but no qualitative changes have been observed in many thousands of additional iterations of the hill-climbing algorithm.

Now what is the optimal distance x for maximising the distinctiveness? As it turns out, the optimal x depends on the noise level δ. Recall that distinctiveness D is defined as the average probability of correct recognition (equation 4.12). In this case, we have three terms describing the recognition probabilities of each of the three signals. These are:

$$P(t_1|t_1) = D_1(x) = \frac{f(0)}{f(0) + f(x) + f(1)} \tag{4.14}$$

$$P(t_2|t_2) = D_2(x) = \frac{f(0)}{f(x) + f(0) + f(1-x)} \tag{4.15}$$

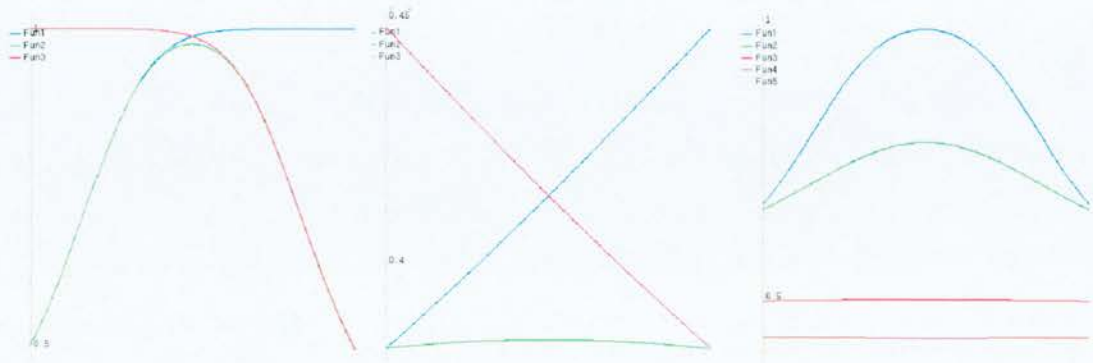
$$P(t_3|t_3) = D_3(x) = \frac{f(0)}{f(1) + f(1-x) + f(0)} \tag{4.16}$$

The values of these three functions, for two different choices of the parameter δ are plotted in figure 4.7(a) and (b). If we add up these three curves, we find, for different values of δ, the curves in figure 4.7(c). Clearly, for low levels of noise the optimal value of x is $x = 0.5$. For higher noise levels, however, this optimum disappears, and the optimal configuration has $x = 0$ or $x = 1$. That is, if there is too much noise, it is better to have several signals overlap. At $\delta = 1.0$ (lowest curve in c), distinctiveness as a function of x is a slightly hollow curve.

Figure 4.8(a) shows a 2-dimensional system of 9 points optimised for distinctiveness with a high noise level ($\delta = 1$). The optimal configuration under these conditions is to have each signal in one of the four corners: 3 corners with 2 signals, and one corner with 3 signals. With this configuration, the distance between the two or three signals that share a corner is $d = 0$, and their confusability high. But at least the distance to the other signals is high ($d = 1$, or $d = \sqrt{2}$).

Maximising distinctiveness is here dominated by maximising summed distance, because the f-scores are alomost linear with distance. Consider a signal in the bottom right corner, and consider it moving to the left, that is, away from the two signals already in that corner. The gain in distance from the bottom-right corner (Δdbr), will be exactly cancelled out by the loss in distance from the bottom left corner (Δdbl). The gain in distance from the top-right corner, however (Δdtr), will not compensate for the loss in distance from the top-left corner (Δdtl). To see why, consider moving the signal a distance ϵ to the left. The summed distance will increase only if the gain in distance to the top-right:

$$\Delta dtr^2 = [\epsilon^2 + 1] - [1] = \epsilon^2, \tag{4.17}$$

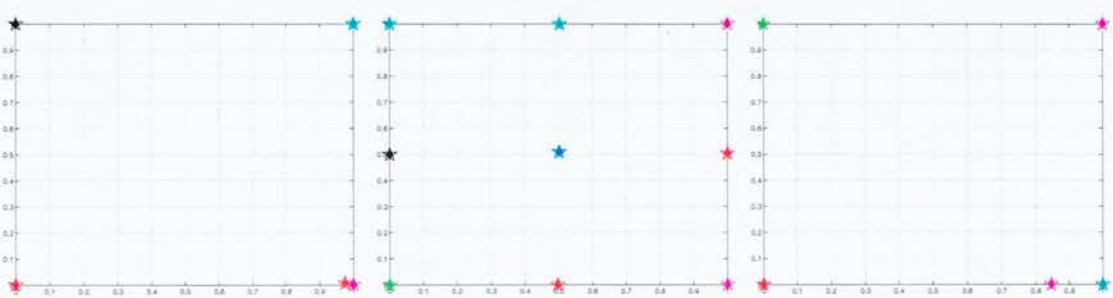


(a) $D_1(x)$ (bottom-left to top-right), $D_2(x)$ (bottom-left to bottom-right) and $D_3(x)$ (top-left to bottom-right) at $\delta = 0.1$.

(b) $D_1(x)$ (bottom-left to top-right), $D_2(x)$ (bottom-left to bottom-right) and $D_3(x)$ (top-left to bottom-right) at $\delta = 1$.

(c) $D(x) = \frac{1}{3}(D_1(x) + D_2(x) + D_3(x))$ at $\delta = 0.1, 0.2, 0.8, 1.0$ (top to bottom).

Figure 4.7: Distinctiveness as dependent on distance and noise, 1d example



(a) 9 points, high noise level ($\delta = 1.0$)

(b) 9 points, low noise level ($\delta = 0.1$)

(c) 40 points, high noise level ($\delta = 1.0$)

Figure 4.8: The noise level determines how many signals can be kept distinct (in a and c the configurations shown are close to convergence, in b it is after convergence).

is larger than the loss in the distance to the top-left:

$$\Delta dtl^2 = [1+1] - [(1-\epsilon)^2 + 1] = [1+1] - [1 - 2\epsilon + \epsilon^2 + 1] = 2\epsilon - \epsilon^2, \tag{4.18}$$

which is never the case if $0 \leq \epsilon \leq 1$.

In contrast, in figure 4.8(b) a system of 9 points is shown that has been optimised for distinctiveness at a relatively low noise level ($\delta = 0.1$). Here maximising distinctiveness is not equivalent to maximising summed distance, because of the relatively low noise level. To see why the noise level determines whether it is equivalent, consider a small change to the configuration, for instance moving the central point a bit to the left. Such a change will decrease the distance to some points, and increase the distance to some other points. Now, note that the distance-to-confusion function is approximately linear for relatively small distances (see figure 4.9). Therefore, maximising distinctiveness corresponds approximately to maximising average distances only if distances are small *relative to the noise level*, or equivalently, if the noise-level is high *relative to the distances.*

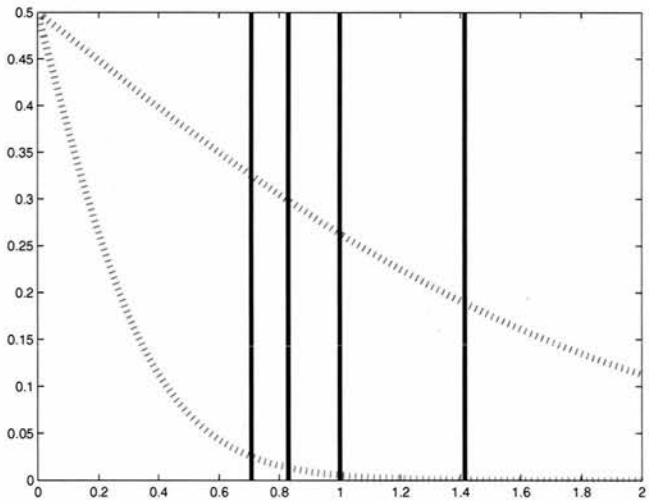


Figure 4.9: The f-scores (y-axis) as a function of distance (x-axis). $\delta = 1.0$ (top curve), $\delta = 0.1$ (bottom curve).

4.4.3 Distinctiveness is a non-linear combination of distances

Figure 4.10 shows another 2-dimensional, 9 signal system. It has, after running the hill-climbing algorithm, converged to a different local optimum (a). Why is this configuration stable? Consider moving the signal α at the right-most end of the line, along that same line. For each alternative x-coordinate of that signal, we can calculate the estimated probability of confusion with other signals. The f-values for all the other signals are plotted in figure (b). For instance, the f-value

of the central-left signal (its contribution to the confusion about α) go from very high (0.3) if α would be on the left-most end of the line, to very low if α is at the right-most end of the line.

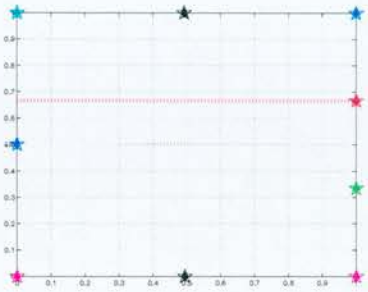
The probability of correct recognition of α, and hence its contribution to the total distinctiveness, is inversely proportional to the sum of all f-values. In figure 4.10(c) I therefore give a plot of the sum of all these values (with the contribution of each signal indicated in different colors). That sum is in a local minimum at the actual location of α, which suggests that – at least initially – distinctiveness will not improve by shifting α to the left. The plot doesn't tell the whole story, though, because the probability of correct recognition of the other signals will also change due to the new position of α. Nevertheless, it does illustrate that the distinctiveness of a repertoire is a non-linear combination of the distances between the signals. Due to this non-linearity, the resulting stable configurations are sometimes counter-intuitive.

4.4.4 Why trajectories stretch out

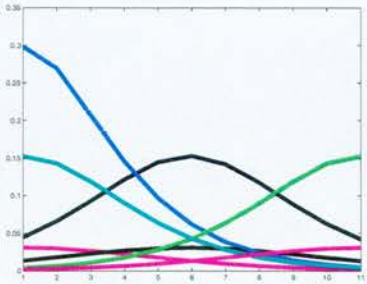
Finally, in figure 4.11 I explore the question of why many trajectories in my simulations stretch out. In figure (a) I show 5 signals (in the bottom-left corner there are 2 signals on top of each other). The signals are points in the acoustic space, which I will here interpret as *stationary* trajectories of some arbitrary length. The graph shows the configuration that maximises the summed distance between the signals. The figure also gives the distance matrix, that gives for every pair of signals the distance between them. Of course, the values are $\sqrt{2} \approx 1.4$ (across the diagonals), 1 (horizontally or vertically) and 0 (for the pair in the bottom-left corner). The average distance is $\overline{d} = 10.2/10 = 1.02$.

Figure 4.11(b) shows an alternative configuration, with the fifth signal in the center. The distance matrix shows that the distance of the fifth signal to the bottom-left corner has increased, but at the expense of the distances to the three other corners. As a result, the average distance has actually gone down to $\overline{d} = 0.96$. The reason is that this configuration doesn't make optimal use of the longest available distances over the diagonal. Importantly, however, at low noise-levels, the distinctiveness of this configuration is in fact higher than of the configuration in (a). The reason is that with relatively little noise and long distances, the distinctiveness–distance function flattens out. Hence, there is more to be gained from avoiding confusion between the fifth and the bottom-left signal, than there is from maintaining the excessive "safety margin" with the other signals. In other words, the configuration in (b) sacrifices some average distance, to gain a more even distribution of distances and, hence, a lower average confusion probability.

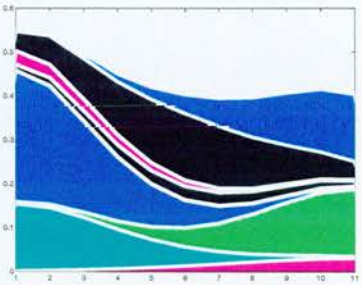
In a restricted space, increasing the distance with one sound will usually decrease the distance with another sound. That is, there is a crucial trade-off between maximising one distance at



(a) 9 points at a local optimum; I will consider moving one signal α along the horizontal cline.

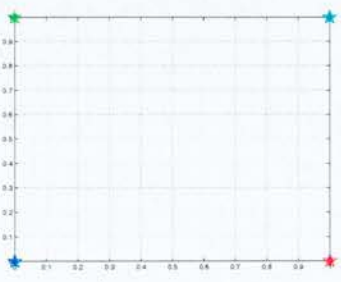


(b) The f-scores of all other signals when α is moved along the cline in (a).



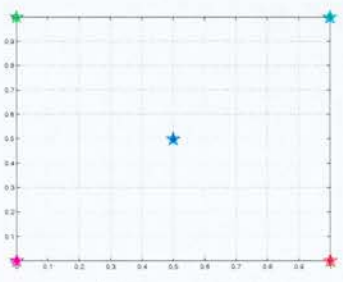
(c) The sum of all the f-scores of all other signals when α is moved along the cline in (a).

Figure 4.10: Figure (a) shows a local optimum of a 9-signal repertoire optimised for distinctiveness. What would happen if we move the signal at the right end of the cline in (a) horizontally to left? The probability of correct recognition of that signal, α, is inversely proportional to the sum of the f-scores of all other signals (see equation 4.11). Figures (b) and (c) show why this probability is in a local optimum with α at its current location. Parameters: $\delta = 0.1$.



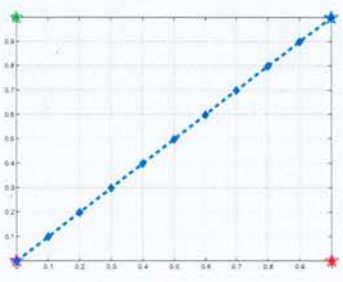
	bl_1	tr	tl	br
tr	1.4			
tl	1	1		
br	1	1	1.4	
bl_2	0	1.4	1	1

(a)



	bl	tr	tl	br
tr	1.4			
tl	1	1		
br	1	1	1.4	
c	0.7	0.7	0.7	0.7

(b)



	bl	tr	tl	br
tr	1.4			
tl	1	1		
br	1	1	1.4	
c	0.7	0.7	0.8	0.8

(c)

Figure 4.11: Why do trajectories stretch out? Three configurations and their distance matrices. Abbreviations: bl: bottom-left, tr: top-right, tl: top-left, br: bottom-right, c centre.

the expense of another. Although maximising distinctiveness D will generally lead to larger distances d, due to the non-linear dependence of D on d, that trade-off can work out differently when maximising D than when maximising $\overline{d}$.

Figure 4.11(c) shows yet another configuration, now with the fifth trajectory stretched out over the whole diagonal. As is clear from the given distance matrix, this configuration yields larger distances than in (b). To go from (b) to (c) there is no trade-off. The distances from the central, fifth signal to the top-left and bottom-left corners can be increased without decreasing the distances to the other 2 signals. The reason is that the distance between a stationary trajectory t and a stretched out trajectory t' is equal to the distance between t and the centroid of t' when t (like the top-right and bottom-left signals) is on a line through all the points of t', but larger when it's not (like the top-left and bottom-right signals). The distinctiveness in (c) is even larger than in (b).

In figure 4.12 I show results from running the basic model under various parameter settings, including with repertoires with many trajectories and with 3-dimensional acoustic spaces. These results show that the observations made in the simple systems above generalise to a wide range of conditions.

4.5 Invasibility

4.5.1 Game-theoretic analysis

So-far, we have seen that repertoires of signals with a temporal structure will, when optimised for distinctiveness, not be organised in as many little clumps as needed, but instead stretch out. Rather than staying away as far as possible from other trajectories along its whole length, each trajectory will be close to some trajectories for some of its length, and close to other trajectories elsewhere. In qualitative terms, these systems show superficially combinatorial phonology. The model represents progress from existing work, because it deals with the discrete and combinatorial aspects as well as with the trade-off between them. It shows a possible sequence of fit intermediates, and, hence, a route up-hill on the fitness landscape.

I have not, however, dealt with the invasibility requirement from chapter 2. Will an innovation, even if it represents an improvement, be able to invade a population where it is very infrequent? To test for invasibility, I adapt the definition of distinctiveness to tell us something about pairs of languages. This way we can ask the question: how well will a repertoire R' do

when communicating with a repertoire R? Pairwise distinctiveness $\mathcal{D}$ is defined as follows:

$$\mathcal{D}(R, R') = \sum_{t=1}^{T} \frac{f(d(R_t, R'_t))}{\sum_{t'=1}^{T} f(d(R_t, R'_{t'}))}. \tag{4.19}$$

The quantity $\mathcal{D}(R, R')$ can be interpreted as the estimated probability of a signal uttered by a speaker with repertoire R, to be correctly interpreted by a hearer with repertoire R'.

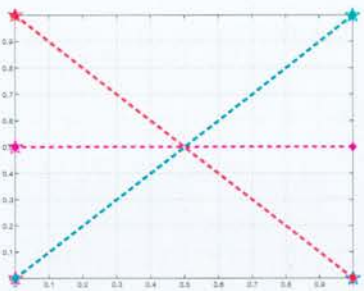
When we now consider the invasion of a *mutant* repertoire R' into a population with *resident* repertoire R, four measures are of interest: $\mathcal{D}(R, R)$, $\mathcal{D}(R, R')$, $\mathcal{D}(R', R)$ and $\mathcal{D}(R', R')$. That is, how well does each of the repertoires fare when communicating with itself or with the other repertoire, in the role of speaker or of hearer? Specifically, for the invasion of R', it is necessary that $\mathcal{D}(R', R) > \mathcal{D}(R, R)$ or $\mathcal{D}(R, R') > \mathcal{D}(R, R)$, or some weighted combination of these requirements (depending on the relative importance of speaking and hearing). That is, a successful mutant must do better against the resident language, than the resident language does against itself. Can such situations arise?

Interestingly, this situation turns out to be very common. Consider the following 1d example:

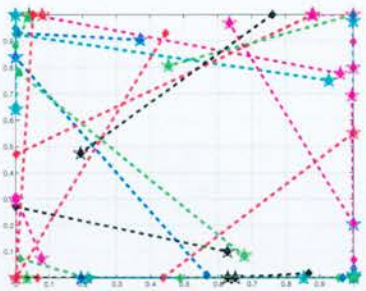
The configuration on the right (B) is better on all accounts. Obviously, there will be less confusion between its signals because they are further apart (when $x = 0.1$ and $\delta = 0.1$, $D(A) = \mathcal{D}(A, A) = 0.70$ vs. $\mathcal{D}(B, B) = 0.84$). But configuration B will even do better when communicating with A, both as a hearer ($\mathcal{D}(A, B) = 0.78$) and as a speaker ($\mathcal{D}(B, A) = 0.76$). The reason is that by having its prototypes more pronounced, f-scores of the wrong signals decrease more than the f-score of the right signal. This is illustrated in figure 4.13 for the slightly exaggerated case of $x = 0.45$. In this example, the f-scores of distances to the left-most signal in A follow a linear regime (a decrease of ~ 0.1 at each step), whereas the f-scores of the distances from B's leftmost signal to the signals in A follow an exponential regime (a decrease of $\sim 50\%$ at each step; see figure 4.13(d)).

Figure 4.14 and 4.15 show results from simulations with improved pairwise distinctiveness as the optimisation criterion. Here, at every step of the hill-climbing algorithm a random change to the existing resident repertoire R was considered, and kept only if the following condition is met:

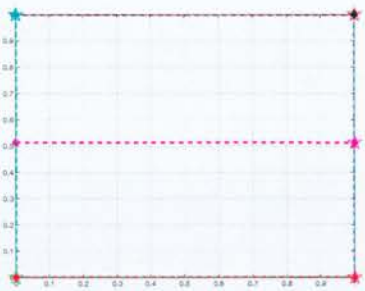
$$\mathcal{D}(R, R') > \mathcal{D}(R, R) \tag{4.20}$$



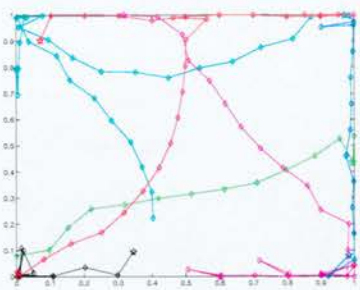
(a) 9 trajectories, length 2, segment size unconstrained



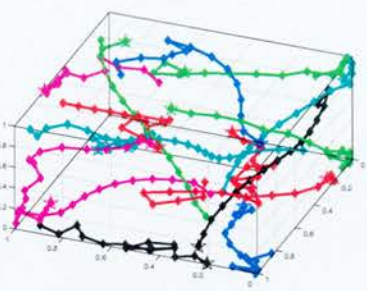
(b) 40 trajectories, length 2, segment size unconstrained



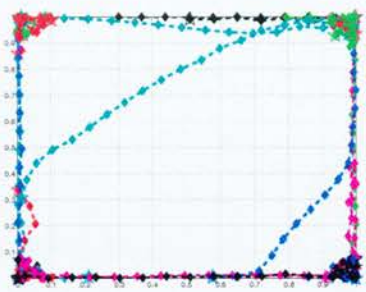
(c) 9 trajectories, length 2, segment size $S = 1.0$



(d) 9 trajectories, length 15, $S = 0.1$



(e) 16 trajectories in 3d, length 15, $S = 0.1$



(f) 30 trajectories, length 20, $S = 0.07$

Figure 4.12: Various signal systems optimised for distinctiveness. Common parameters: $\delta = 0.1$, $\rho = 0.05$.

$$\mathbf{d_{A,A}} = \begin{pmatrix} 0 & .05 & .1 \\ .05 & 0 & .05 \\ .1 & .05 & 0 \end{pmatrix}$$

(a) distances A vs. A

$$\mathbf{d_{B,A}} = \begin{pmatrix} .45 & .5 & .55 \\ .05 & 0 & .05 \\ .55 & .5 & .45 \end{pmatrix}$$

(b) distances B vs. A

$$\mathbf{d_{A,B}} = \begin{pmatrix} .45 & .05 & .55 \\ .5 & 0 & .5 \\ .55 & .05 & .45 \end{pmatrix}$$

(c) distances A vs. B

d	$f(d)$	d	$f(d)$
0.0	0.500	0.45	0.012
0.05	0.401	0.5	0.006
0.1	0.309	0.55	0.003

(d) f-scores

$$\mathbf{U_{A,A}} = \begin{pmatrix} .41 & .33 & .26 \\ .31 & .38 & .31 \\ .26 & .33 & .41 \end{pmatrix}$$

(e) confusion A vs. A

$$\mathbf{U_{B,A}} = \begin{pmatrix} .57 & .29 & .14 \\ .31 & .38 & .31 \\ .14 & .29 & .57 \end{pmatrix}$$

(f) confusion B vs. A

$$\mathbf{U_{A,B}} = \begin{pmatrix} .57 & .29 & .14 \\ .01 & .98 & .01 \\ .14 & .29 & .57 \end{pmatrix}$$

(g) confusion A vs. B

Figure 4.13: Distance matrices, f-scores and confusion matrices. Parameters: $\delta = 0.1$, $x = 0.45$.

Hence, the algorithm is almost identical to the hill-climbing algorithm used before, but with the criterion of *distinctiveness* replaced by the criterion of *pairwise distinctiveness*:

```
% R is a repertoire of signals
% S is the segment length parameter
% ρ is the hill-climbing rate parameter
% δ is the acoustic noise parameter (characteristic distance)
for i = 1 to I
    R' = CONSTRAIN( R + DISTURBANCE( ρ ), S);
    if D(R,R',δ) > D(R,R,δ) then R = R';
end for
```

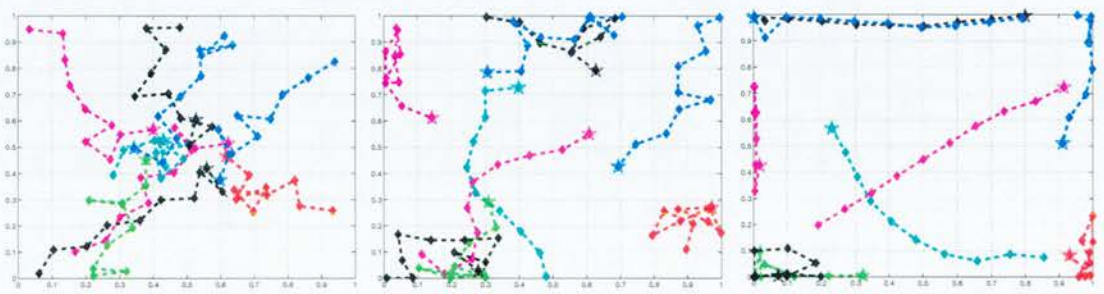
Figures 4.14(a-f) show the configuration of the repertoire at different numbers of iterations of the hill-climbing algorithm. Figure 4.14(g) gives the pairwise distinctiveness measures for each combination of these 6 configurations. At the diagonal of this matrix are the distinctiveness scores of each configuration. As is clear from this matrix, each next configuration can invade a population with the previous repertoire. In bold-face we see the approximate evolutionary trajectory (the actual steps in the simulation are much smaller and much more numerous). As is clear from figure 4.14(f), the final configuration shows the same type of superficially combinatorial phonology that I found in the straightforward optimisation version of the model.

This final configuration is probably an Evolutionarily Stable Strategy, as defined in chapter 2. However, the condition of equation (4.20) is slightly weaker than the condition for an ESS. To establish rigorously that this configuration R is an ESS we would need to show there is no alternative configuration R' where $\mathcal{D}(R,R') \geq \mathcal{D}(R,R)$, or if there is such a configuration that $\mathcal{D}(R,R) > \mathcal{D}(R',R')$. The condition here also differs from the condition used in Nowak & Krakauer (1999). Whereas these authors assume that the total payoff is the exact average of payoff as a hearer and payoff as a speaker (equation 4.3), in these simulations only the payoff as a hearer was modelled[9]. I expect the behaviour of the model to change very little, but it would be worthwhile to investigate the behaviour of the model with these different optimisation criteria. These simulations have not yet been performed.

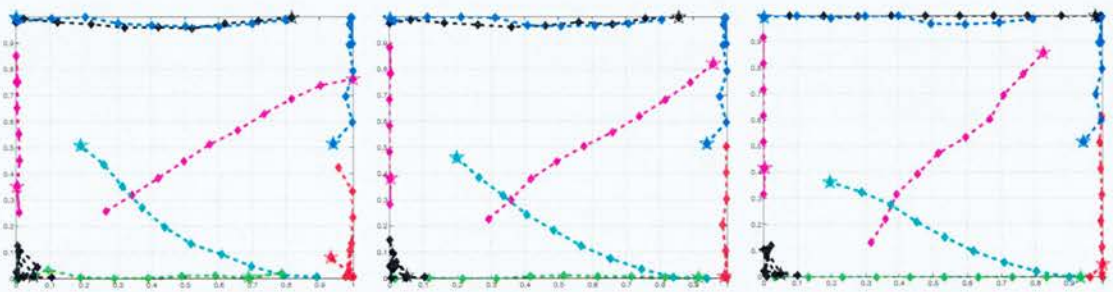
4.5.2 *Individual-based model*

As a final test of the appropriateness of the basic model, Bart de Boer and I studied an individual-based simulation of a *population* of agents that each try to imitate each other in noisy conditions.

[9]That is, I used $\mathcal{D}(R,R') > \mathcal{D}(R,R)$ rather than $(\frac{1}{2}\mathcal{D}(R,R') + \frac{1}{2}\mathcal{D}(R',R)) \geq \mathcal{D}(R,R)$. Thanks to Matina Donaldson for pointing this out.



(a) t=0 (b) t=500 (c) t=2000



(d) t=4000 (e) t=8000 (f) t=13000

$$\mathbf{D}^* = \left(\begin{array}{c|cccccc} & a & b & c & d & e & f \\ \hline a & \mathbf{.24} & \mathbf{.34} & .33 & .33 & .33 & .33 \\ b & .37 & \mathbf{.50} & \mathbf{.50} & .49 & .49 & .49 \\ c & .41 & .51 & \mathbf{.71} & \mathbf{.70} & .69 & .68 \\ d & .39 & .52 & .71 & \mathbf{.76} & \mathbf{.76} & .75 \\ e & .39 & .52 & .71 & .76 & \mathbf{.77} & \mathbf{.77} \\ f & .39 & .52 & .70 & .75 & .77 & \mathbf{.79} \end{array}\right)$$

(g) The pairwise distinctiveness matrix

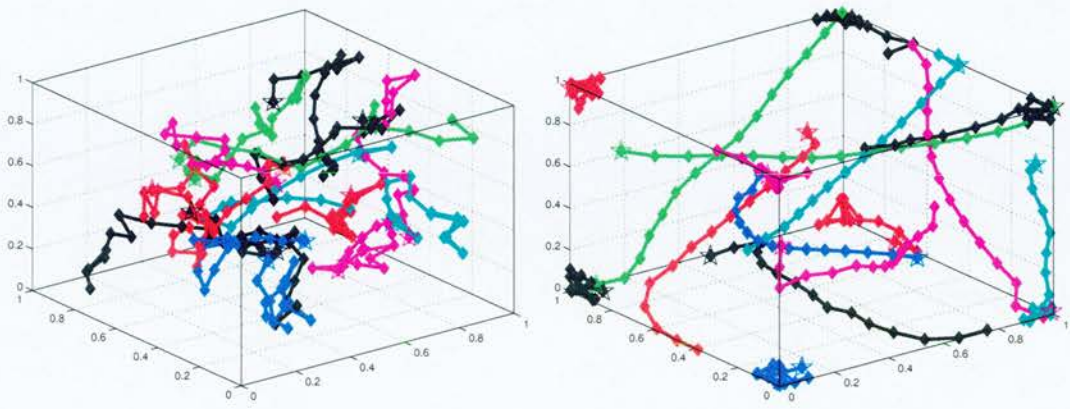
Figure 4.14: Invasability experiments. Parameters are: T=9, P=10, D=2, N=0.05, S=0.1. The confusion-probabilities are proportional to $\frac{1}{1+e^{(\frac{1}{\delta}d^2)}}$, where $\delta = 0.1$ and d is the average segment-by-segment Euclidean distance. The approximate evolutionary trajectory is indicated with boldface in the pairwise distinctiveness matrix of figure (g).

This simulation (the current version implemented by Bart de Boer) is similar to the model described above, but now each agent in the population has its own repertoire, and it tries to optimise its own success in imitating and being imitated by other agents of the population. Hence, a random change, as before, is applied to a random trajectory of a random agent in the population. If this change improves the imitation success in interaction with a number of randomly chosen other individuals in the population, it is kept. Otherwise, it is discarded.

This version of the model is like the imitation games of de Boer (2000). These only modelled holistic signals (vowels) and did not investigate combinatorial phonology. The game implemented here is a slight simplification of the original imitation game. First, all agents in the population are initialised with a random set of a fixed number of trajectories. Then for each game, a speaker is randomly selected from the population. This speaker selects a trajectory, and makes a random modification to it. Then it plays a number of imitation games (50 in all simulations reported here) with all other agents in the population. In these games, the *initiator* utters the modified trajectory with additional noise. The *imitator* finds the closest trajectory in its repertoire (according to the Dynamic Time Warping, DTW, distance metric) and utters it with noise. Games are successful if the imitator's signals is closest to the modified trajectory in the initiator's repertoire. If it turns out that the modified trajectory has better imitation success than the original trajectory, the modified trajectory is kept, otherwise the original one is restored.

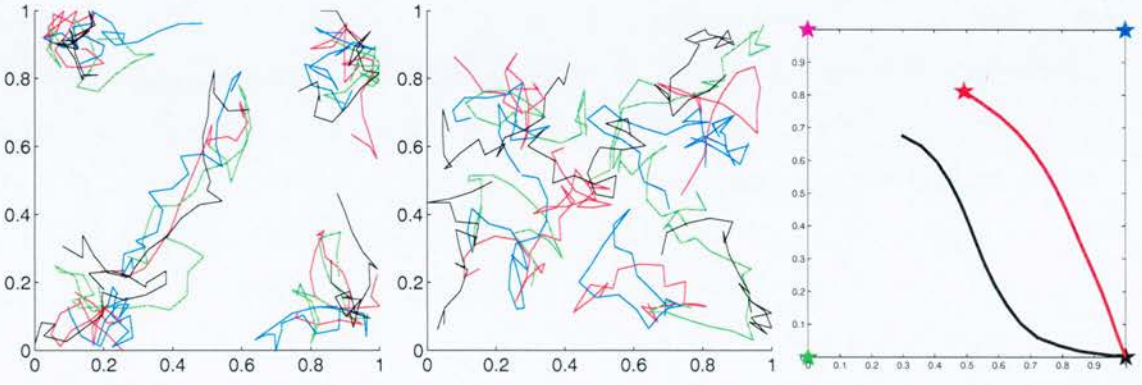
For vowel systems, it has been shown that optimising a single repertoire leads to similar systems as a population-optimisation system (compare Liljencrants & Lindblom, 1972; de Boer, 2000). As it turns out, also for repertoires of trajectories these two types of models yield comparable results, at least if the noise on the trajectories is time-correlated. That is, if distortions of a point on the trajectory are not completely independent from distortions of its neighbouring points. This is illustrated in figure 4.16.

In this figure the left frame shows the system of five trajectories that resulted from playing imitation games in a population, using form-preserving noise. The right frame, for reference, shows a system of five trajectories that resulted from optimising total distance (DTW metric) as in the basic model. Observe that in both cases, the corners are populated by four trajectories, which are bunched up. The fifth trajectory, in contrast, follows the diagonal. As before, an analysis in terms of phonemes suggests itself: the four corners are basic phonemes, while the fifth trajectory uses one as the corners as a starting phoneme and the opposite corner as the ending phoneme. Both models result in similar systems of trajectories.



(a) t=1000 (b) t=11000

Figure 4.15: A 16 trajectories signal system in a 3d acoustic space, after 1000 and 11000 iterations. At each time step, a small random change is considered, and only adopted if it represents an improvement according to the pairwise distinctiveness criterion (equation 4.19). Parameters are: $P = 15, S = 0.1, \rho = 0.05, \delta = 0.1$. The final distinctiveness is $D(R) = \mathcal{D}(R, R) = 0.94$.



(a) (b) (c)

Figure 4.16: Comparison of population-based models with the optimisation model. Frame (a) shows the (five) trajectories of four agents (from a population of ten), when noise preserves the shape of trajectories. Notice the similarities with the optimised trajectories in frame (c). If noise does not preserve shape of trajectories, the trajectories tend not to stretch out, as shown in the frame (b). Although it is rather hard to see, there are four clusters in the corners, and one in the middle. (Graphs created by Bart de Boer).

The middle frame, on the other hand, shows that when noise is not time-correlated, a system results in which all trajectories are bunched up and an analysis in terms of phonemes is therefore not possible. As noise in real signals is band limited, it follows that shape will always be preserved to some extent. For computational reasons, we have not performed simulations in the population condition with more than 5 trajectories. However, although less clean and not fully conclusive, the results from the individual-based model seem to be consistent with the observations with the basic model.

4.6 Conclusions

Natural language phonology is discrete and combinatorial. What I have shown in this chapter is that these properties have functional significance: they aid the reliable recognition of signals by the hearer. I have also shown that there is a series of steps that lead from a signal system without to a system with these properties. Crucially, I have shown that each step in this series represents an improvement, both in populations where it is extremely rare (invasibility), and in populations where it is common.

These findings are consistent with several rather different scenarios for the emergence of combinatorial phonology in the human species. In the simplest, but perhaps least plausible scenario, the starting point is an ancestral population of individuals, each with a heritable repertoire of holistic signals. At some point, a genetic mutation occurred in one individual that changed her repertoire slightly toward combinatorial phonology. With the mutation, this individual communicated slightly more successfully with the others in the population, even though her repertoire was not identical to theirs. Consequently, the new genotype conferred a slightly higher fitness, and the mutated gene spread in the population. Then a second mutation occurred, and a third etc. Many such slight modifications eventually led to a situation where the language of the population was superficially combinatorial, such that the capacity for productive combination, another type of gene mutation, could invade.

The problem with this scenario is that it is inconsistent with the facts about the acquisition of phonology, at least in modern humans. Languages differ enormously in the number and nature of the units of combination, but there seems to be no heritable variation in the ability to acquire any of these many different phonologies. Perhaps this interpretation of my model has more relevance when applied to combinatorial vocalisations in other species, but for human phonology it is unrealistic to postulate many genes responsible for the specific shape of the sounds in our languages.

An alternative scenario is that combinatorial phonology arose in a population of learners that each optimise their success in recognising and being recognised by making small adaptations to their repertoire of signals. Combinatorial phonology, in this scenario, is the result of the interaction between many individuals in a population – a process that could be called "self-organisation".

The problem with such a scenario is that it requires powerful learning abilities of individuals, such that, on average, the adaptations made represent improvements. That is, an individual must be able to change her repertoire and accurately assess the relative benefits of the change in the communication with others in the population. Modern humans seem to be able to make such adaptations, both in first and in second language acquisition. But where does this successful learning ability come from? Perhaps it was just there, by lucky coincidence, as a side-effect of the evolution of other cognitive abilities; or perhaps the learning ability itself evolved by natural selection. However, by simply postulating that there is such an ability, we have shifted the explanatory challenge from the emergence of repertoires of signals, to the emergence of learning procedures for repertoires of signals.

I argue that there is a third scenario that represents a middle way, and avoids both the excessive genetic determinism of the first scenario, and the reliance on happy coincidences of the second scenario. Natural selection favours fitter individuals over less fit ones, but it is, in a sense, blind for the (proximate) causes of the fitness differences. All other things being equal, natural selection cannot tell the difference between an individual born with an "innate" repertoire of signals R, and an individual that learned that same repertoire from experience. Hence, a mutation that changes a learning rule such that it leads to a slightly more distinctive repertoire than that of the resident population, is favoured by natural selection under the same conditions as any other mechanism with the same effect (Harley, 1981; Maynard Smith, 1982, chapter 4).

We can thus view the signals in the model of this chapter as the outcome of a process of learning from signals used in the population. The resident learning strategy will learn the repertoire as is[10]; the mutant learning strategy will learn a different repertoire. The mutant will be favoured by selection only if the differences are slight, and if the repertoire learned is more distinctive. Once adopted by a significant fraction of the population, the mutant learning strategy will itself change the shape of the population's repertoire: a process of self-organisation kicks in and makes the repertoire more combinatorial.

In this view, self-organisation and natural selection are not *alternative* explanations for the phenomenon of combinatorial phonology. Rather, natural selection shapes the parameters of the

[10]The repertoire will reflect the resident learning strategy employed in previous generations. See chapter 7.

self-organising process. Hence, *self-organisation is the substrate of evolution* (Thompson, 1932; Waddington, 1939; Boerlijst & Hogeweg, 1991). With such an interpretation, the model of this chapter is consistent with the ideas on selforganisation of Lindblom *et al.* (1984), de Boer (2001) and Oudeyer (2002), while solving some of the problems with their formal models, as well as with those of Liljencrants & Lindblom (1972) and Nowak & Krakauer (1999).

CHAPTER 5

Compositional Semantics[1]

Compositional semantics – where the meaning of a combination is a function of the meanings of the parts – is a fundamental property of natural language. Explaining its evolution remains a challenging problem because existing explanations require a structured language to be present before compositionality can spread in the population. In this chapter, I study whether a communication system can evolve that shows the preservation of topology between meaning-space and signal-space, without assuming that individuals have any prior processing mechanism for compositionality. I present a formalism to describe a communication system where there is noise in signalling and variation in the values (payoffs) of meanings. In contrast to previous models, both the noise and payoffs depend on the topology of the signal- and meaning spaces. I consider a population of agents that each try to optimise their communicative success. The results show that the preservation of topology follows naturally from the assumptions on noise, payoffs and individual-based optimisation.

[1]This chapter describes research that builds on joint work with Gert Westermann, as published in Zuidema & Westermann, 2003 (see appendix C of this thesis). However, all modelling, text and graphs in this chapter are my own. Some results have been published in Zuidema, 2003c (see appendix C).

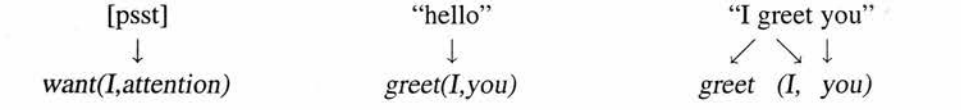
5.1 Compositionality in Natural Language

After discussing the evolution of combinatorial phonology in the previous chapter, I will now focus on the evolution of another combinatorial principle that is a universal property of natural language: *compositional semantics*, or "compositionality". It can be defined as follows:

Definition 2 (Compositionality) *A language is compositional if it consists of a set of meaningful units which can be combined into larger wholes, in such a way that the meaning of the whole is a function of the meaning of the parts and the way they are put together.*

From a formal point of view, this definition is more ambiguous than one would like, because the words "meaning" and "function" leave much room for interpretation depending on the theoretical framweork one works in. In the philosophy of language, theoretical linguistics and mathematical logic a vast literature is associated with "compositionality" (see Janssen, 1997 for a review), and a strict interpretation of the concept exists that excludes many of the example construction I will give below. In this thesis, however, I will interpret compositionality in a broad, intuitive sense, as is common in the field of language evolution (e.g. Batali, 1998; Wray, 1998; Kirby, 2000; Brighton, 2002; Hurford, 2002a). In this interpretation, compositional semantics is in contrast with a *holistic* semantics, where the meaning of an utterance cannot be derived from the meaning of its parts at all.

However, I will restrict the use of the word "compositionality" to refer to a property of the mapping between forms and meanings. That is, in my terminology, "meanings" cannot be compositional; they can be *combinatorial*, i.e. be built-up from units that can be combined in many different ways[2]. Similarly, forms are never compositional, but might be combinatorial. Thus, utterances with *holistic semantics* might show *combinatorial phonology*, that is they can be built-up from *meaningless* elements (phonemes, syllables). The use of the words "holistic", "combinatorial" and "compositional" in this thesis, is illustrated in figure 5.1.



(a) Holistic phonology, holistic semantics

(b) Combinatorial phonology, holistic semantics

(c) Combinatorial phonology, compositional semantics

Figure 5.1: Holistic, combinatorial and compositional

[2]At several points in this thesis, I will use (predicate) logic notation for such meanings. For logicians, such expressions have a semantics themselves (for instance, its truth conditions given a model) which can be compositional. In this thesis, however, I will only be concerned with the mapping between (phonological) form and (conceptual) meaning.

Together, combinatorial phonology and compositional semantics constitute **duality of patterning** (Hockett, 1960). In English, compositionality in its broad sense is evident in, for instance, productive morphology (such as the plural *-s* and the regular past tense suffix *-ed*), compound noun constructions (such as *"thesis deadline"*, *"dog house"* and *"bear country"*), compound verbs (*"move on"*, *"move out"*, *"keep on"*, *"keep out"*) and phrasal syntax (*"John sees Mary"*, *"Mary sees John"*, *"Did Harry ever meet Sally?"*). Natural languages use a range of compositional mechanisms, with much variation between languages in which mechanisms are used, and to which degree. English, for instance, makes, in comparison with Turkish or Russian, very little use of productive morphology.

As in phonology, the units of combination in each of these compositional mechanisms remain a topic of debate. In morphology, it has long been recognised that utterances that look as if they are composed of smaller units, might in fact be stored and retrieved as whole chunks. This is referred to with the term *semi-productive morphology*. For instance, in what language typologists call *fusional languages*, one can identify word stems and suffixes, but the combinations are not completely regular, i.e. there is no clear-cut boundary between the morphemes in a word. Rather, the morphemes are often "fused" together and give a single, unsegmentable morph. An example is Russian, where the words *stol* for "table" and *lipa* for "lime-tree" are inflected as follows (examples from Comrie, 1981):

	singular I	plural I	singular II	plural II
nominative	*stol*	*stol-y*	*lip-a*	*lip-y*
accusative	*stol*	*stol-y*	*lip-u*	*lip-y*
genitive	*stol-a*	*stol-ov*	*lip-y*	*lip*
dative	*stol-u*	*stol-am*	*lipea*	*lip-am*
instrumental	*stol-om*	*stol-ami*	*lip-oj*	*lip-ami*
prepositional	*stol-e*	*stol-ax*	*lip-e*	*lip-ax*

Such imperfect compositionality poses a challenge to formal models of language. On the one hand, the irregularities force one to store the larger chunks as unanalysed wholes in a lexicon. In the Russian example above, a formal model would need to store the genitive singular form of *"lipa"* as one unit *"lipea"*, because that form cannot (or at least, not obviously) be derived from the stem *"lip"* and the regular suffix *-u* or *-a*. On the other hand, one would like a formalism to take advantage of the manifest regularities in the system. For English, this tension has led to the on-going *past tense debate*. In this debate (see e.g. Clahsen, 1999, and the peer-commentaries in

the same issue), one side denies the cognitive relevance of semi-regular inflections such as *sing-sang-sung*, *ring-rang-rung*, and views the regularities solely as remnants of an older stage of the English language. The other side argues that semi-regular and regular inflection is all part of the same system, and that the right model of past tense inflection should therefore include associative mechanisms.

More recently, researchers from diverse subfields of linguistics and from many different schools, have argued that similar phenomena exists at all levels of language processing, from phonology to syntax[3]. For instance, Bybee (2003) has argued that the construction *"it's"*, although obviously a contraction of *"it is"*, has linguistically and psychologically the profile of a single word. In usage-based theories of language acquisition, it is proposed that children first acquire large chunks and only later discover the words and morphemes they consist of (Langacker, 1987; Tomasello, 2000; Lieven *et al.*, 2003). In computational linguistics, Bod (1998, 2003) has shown that large-coverage parsers benefit from storing large chunks of parse trees from a tree bank. Finally, even within generative linguistics, the issue of the unit of storage and the need for a heterogeneous model has recently come to the forefront (Nooteboom *et al.*, 2002; Jackendoff, 2002).

For theories of the evolution of compositional semantics, the issue of the unit of combination is relevant in two ways. First of all, it matters of course what exactly needs to be explained. If, counter-factually, morphology were not productive at all, there would be no need for an evolutionary explanation for productive compositionality at this level of natural language. Second, the existence of semi-productive, or "superficial" compositional patterns in languages raises the question of where these patterns come from. It seems that either they are the remnants of an earlier phase of the language (which implies that productive compositionality can become non-productive in historical language change), or they are the result of other mechanisms which bring about the appearance of compositionality. Both routes to superficial compositionality suggest an interaction between storage and productive mechanisms, and hold important clues for theories of the evolutionary origin of productive compositionality.

5.2 The Evolution of Compositionality

Although the unit of combination in productive morphology and syntax is an open problem, there is consensus that productive compositionality, at some level, is an essential feature of human

[3]From an evolutionary point of view, a mixture of regular and irregular systems is exactly what one would expect. For instance, the evolved "genetic code" that defines the mapping from genes to proteins also combines regular with highly irregular rules (Nick Barton, p.c.; Maynard Smith & Szathmáry, 1995). Models of the cultural evolution of language predict that frequent meanings will be expressed with irregular words, whilst infrequent meanings are expressed with regular combinations (Kirby, 2001). I'll briefly come back to these issues in chapter 7.

languages. In animal signal systems, in contrast, compositionality occurs only rarely and only to a very limited extent. A classic example of a compositional signalling system is the bee dance (von Frisch, 1965, 1974), used to communicate the location of a newly discovered food resource. In these dances two features of the "form", the length and direction of the longest stretch, map on two aspects of the meaning: distance and direction of the food source. Although a fascinating example, the facts that bees are a phylogenetically extremely distant species from humans, that the dances can only be used to communicate distance and direction of food, that the signals are analogue, and that a completely different medium is used, make it of little relevance for the evolution of human language.

In non-human primates, little evidence of spontaneous use of compositionality exists. Chimpanzees and bonobos seem to be capable, typically only after intense training, of using combinations of signs to express compound meanings (Savage-Rumbaugh *et al.*, 1986; Savage-Rumbaugh & Lewin, 1994; Premack, 1971), but the evidence remains disputed (Pinker, 1994). Only anecdotal evidence about compositional communication in wild chimpanzees and gibbons exists (Ujhelyi, 1996; Savage-Rumbaugh, 2000). An intriguing example of compositionality in wild Campbell monkeys has recently been described by Zuberbühler (2002). Like Vervet monkeys (Seyfarth *et al.*, 1980), Campbell monkeys have an alarm call system, with specific calls for a small number of predator categories. In addition, Campbell monkeys have a distinct grunt that modifies the meaning of the call that follows: it weakens the meaning of the following call. In playback experiments with monkeys from an other species (Diana monkeys), Zuberbühler was able to demonstrate that they respond reliably with the appropriate predator response when presented with 3 types of alarm calls, but respond halfheartedly or not at all when these same calls were preceded by the cancel grunt[4].

How did limited compositional communication like in Campbell monkeys and the extensive compositionality of human language evolve? Jackendoff (1999, 2002) includes compositionality ("use of symbol position to convey basic semantic relations") as one of the major stages in his scenario for the evolution of language. In this scenario, productive combination of signs emerged before the systems for word order, phrase-structure, agreement and inflection were in place (see chapter 3). Jackendoff argues that modern languages contain "fossils" of the compositionality stage. For example, the compositional compound noun construction in English mentioned above

[4]It is unclear to me whether the call system is an example of *productive* compositionality; it seems this would be difficult to test, because the alarm calls in monkeys tend to be innate (Seyfarth & Cheney, 1997). If new calls can be taught, one can easily distinguish between a productive system (where the effect of the cancel grunt would have to generalise to the newly taught signal), and a holistic system that is superficially compositional. But if a repertoire of calls is fixed, it is difficult to make this distinction.

can be viewed as such a fossil: the meaning of a compounds like *"gun shot"* and *"shot gun"* is deducible (but not completely specified) from the meaning of the component words and the order in which they are put. That is, typically the second noun in a compound determines its type, making *gun shot* a type of shot (or wound) and *shot gun* a type of gun. The rules are not strict however; *pickpockets* are not a type of pockets, and whereas a *snow man* is made of snow, a *fire man* is not made of fire.

A compositional language, without all the niceties of modern morphosyntax (in particular, hierarchical structure), corresponds roughly to Bickerton's concept of **protolanguage** (Bickerton, 1990). Recall from chapter 3 that protolanguage is the hypothetical precursor of modern language that is assumed to share many characteristics with pidgin languages (the limited languages acquired by adults that need to communicate in a population where there is no dominant language) and "Basic Variety" (the limited proficiency attained by adults learning a foreign language, Klein & Perdue, 1997).

For those researchers who believe in a gradualist scenario for the evolution of human language, the assumption of a protolanguage with limited compositionality as an intermediate stage is relatively uncontroversial. However, the exact properties of protolanguage and its precursors are a topic of a debate (e.g. Wray, 2000; Tallerman, 2005). For instance, Jackendoff imagines a transition from free concatenation to more fixed word order, but Bickerton explicitly rejects this view (Bickerton, 2003a). Given the entirely hypothetical status of "protolanguage", this debate is conducted with surprising vigour. In the end, as I have argued in chapters 1, 2 and 3, only a complete and formal scenario, with convincing explanations for the transitions for one step to the next, will resolve these disputes. Hence, in our efforts to formalise Jackendoff's scenario, the real issue is to explain how the transition from a non-compositional stage to a compositional stage could have happened.

Hurford (2002a) classifies explanations for the transition according to whether they postulate an **analytic route**, where holistic signals are reanalysed as consisting of meaningful parts, or a **synthetic route**, where pre-existing signals are combined in novel combinations. Whereas Jackendoff clearly favours a synthetic route, the model of this chapter will explore an analytic route. As I will argue below, no convincing formal model of an synthetic route has been proposed yet.

As with the other transitions, explanations for the transition to compositionality can be further classified according to their reliance on **language-specific, biological adaptations**, versus their reliance on domain-general learning mechanisms and self-organisation. Many researchers,

including Pinker & Bloom (1990); Nowak & Krakauer (1999); Nowak *et al.* (2000); Jackendoff (2002), have argued for innate, language-specific cognitive specialisations for compositionality that have evolved under natural selection. Such explanations can be further classified according to the assumed function of compositionality. Pinker and Jackendoff's verbal treatments keep this crucial issue rather vague, and refer to a diverse array of advantages, from those for learning, memorising and generalisation, to those for sharing information and for impressing peers and sexual partners. The mathematical models of Nowak and colleagues are more precise, but whereas in the model of Nowak & Krakauer (1999) the selection pressure is acoustic distinctiveness, in Nowak *et al.* (2000) it is rather "learnability". These models are discussed in more detail in the next section. Unfortunately, explanations of this type – even the mathematical models – have generally remained much *underspecified*, and have not adequately dealt with the invasibility constraint and the problems of cooperation and coordination that I discussed in chapter 2. In particular, as I will show below, the only existing formal models of the biological evolution of compositionality, only establish that it leads to more successfull communication *once it has been adopted by the whole population.*

The crucial question is, of course, how much the hominid brain has had to change to be able to process compositional language (Lewontin, 1990). Most researchers, including Jackendoff, agree that prelinguistic hominids, like modern Great Apes, must be assumed to have had a sophisticated, *combinatorial* conceptual apparatus[5]. It is unclear to what extent these hominids could have made use of such pre-existing cognitive abilities for processing a simple compositional language. Although most theories of grammar postulate computational procedures that appear very specific for language, some linguists have argued that the disparity is more apparent than real. For instance, Steedman (2002b) finds that combinatory categorial grammar (a formalism for describing natural language syntax; Steedman & Baldridge, 2003) and classical planning algorithms (from artificial intelligence) make use of exactly the same type of fundamental logical combinators. If the computational requirements for planning, tool use and navigation are not very different from those for processing compositional language, then postulating biological innovations for compositionality might not be necessary. The apparent failure of non-human primates to understand and produce compositional messages, is perhaps better explained by their lack of attention to the intentions of communication partners (Dunbar, 1998; Worden, 1998; Tomasello, 2003).

[5]As I argued in chapter 3, it is difficult to give an exact characterisation of the kind of representations non-human primates, prelinguistic hominids and prelinguistic human infants might have available. For the purposes of this chapter, it is probably safe to think of these representations as something similar to (first-order) predicate logic. This is the assumption made in many formal models of language acquisition (e.g. Pinker, 1979; Wolff, 1982; Buszkowski & Penn, 1990), as well as in most models of the evolution of compositionality (Batali, 1998; Kirby, 2000; Hurford, 2000).

A number of researchers have studied formal models of the transition to compositionality, assuming just general learning and cognitive abilities and **cultural evolution** as the driving force (Batali, 1998, 2002; Kirby, 2000, 2001; Kirby & Hurford, 2002; Hurford, 2000). These models are interesting, and will play a major role in this thesis, but they face some new difficulties themselves as well: (i) in many cases, the assumed cognitive abilities are much more language-specific than one would like; (ii) cultural evolution, such as the progressively better structured languages in the "Iterated Learning Model" (Kirby, 2000; Brighton, 2002), only takes off when there is already some initial structure in the language.

Explaining the evolution of compositionality thus remains a challenging problem because both the genetic and the cultural evolution explanation require a structured language to be already present in the population before the linguistic innovations can successfully spread in a population. In the next section I will first review a number of existing formal models, discuss the mentioned problems in more detail and show that none of the models is complete or conclusive. I will conclude from this review that the need to explore alternative explanations for the evolution of compositional semantics remains.

In the rest of the chapter, I will (unfortunately) not solve this problem. In fact, I will not study the evolution of productive compositionality itself, but focus on superficial compositionality instead. I will start from the observation that compositionality is a form of "topology preservation" in the mapping from meanings to signals. That is, compositionality implies that similar meanings are expressed with similar signals[6].

In line with the consensus, I assume that the meaning space of early hominids was structured, i.e. that the meanings expressed were not holistic, idiosyncratic, categorical objects, as in models of the evolution of a simple lexica (such as signalling games in the tradition of Lewis, 1969, discussed in chapter 2, section 2.6). Moreover, I will assume that communication was noisy, and that very similar signals were easier confused than very distinct signals (as explored in chapter 4). Based on these observations, I present the simplest possible extension of existing models, that allows the similarities between meanings, the similarities between signals and the preservation of topology to be formally described. I will show that in this model optimisation for noise robustness automatically leads to topology preservation. For now, I only look at low dimensional signal and meaning spaces. As a model for the evolution of compositionality it is obviously incomplete, but

[6]There are many difficulties with this statement, because there are many ways to define topologies for meanings and signals, and many ways to define compositionality. For now, I will ignore these problems, but I will briefly come back to this point in the discussion. In the model of this chapter I will consider only simple, one- or two-dimensional Euclidean similarity metrics. I will leave more interesting topologies for future work.

I claim the model does make a start with an alternative route, and suggests possible origins of structure in lexical communication.

5.3 Formal Models of the Evolution of Compositionality

5.3.1 Natural Selection for Compositional Semantics

The first game-theoretic model of the evolution of compositionality was studied by Nowak & Krakauer (1999). The analysis presented by these authors is closely related to their analysis of the evolution of combinatorial phonology, that I discussed in the previous chapter. It is worth considering the applicability of the model to the issue of compositional semantics, and the problems with it, because it is the only formal model that, like the model I develop in this chapter, considers the relation between noise robustness and compositionality.

To briefly recapitulate, Nowak & Krakauer consider a set of available meanings, a set of available signals and three matrices $\mathbf{S}$, $\mathbf{R}$ and $\mathbf{U}$ that describe production, interpretation and confusion respectively[7]. For simplicity, they imagine a world where there are just *objects* and *actions*, and consider two types of strategies, holistic (with a unique word for every object–action combination) and compositional (with nouns for objects and verbs for actions). Their goal is to show that compositionality can evolve, and to identify the conditions under which this will happen. Like in the analysis of combinatorial phonology ("word formation"), Nowak & Krakauer view compositionality ("basic grammatical rules") as a strategy for improving the robustness against noise. They use the same measure for the quality (fitness) of a language as before (equation 4.3):

$$F(L, L') = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathbf{S}_{mi} \mathbf{U}_{ij} \mathbf{R'}_{jm} + \mathbf{S'}_{mi} \mathbf{U}_{ij} \mathbf{R}_{jm} \right). \tag{5.1}$$

The analysis then starts with $\mathbf{S}$- and $\mathbf{R}$-matrices of the following form:

$$\mathbf{S} = \left(\begin{array}{c|cccccccc} & w_1 & w_2 & w_3 & w_4 & N_1V_1 & N_1V_2 & N_2V_1 & N_2V_2 \\ \hline O_1A_1 & 1-x & 0 & 0 & 0 & x & 0 & 0 & 0 \\ O_1A_2 & 0 & 1-x & 0 & 0 & 0 & x & 0 & 0 \\ O_2A_1 & 0 & 0 & 1-x & 0 & 0 & 0 & x & 0 \\ O_2A_2 & 0 & 0 & 0 & 1-x & 0 & 0 & 0 & x \end{array} \right),$$

[7] In this paper (Nowak & Krakauer, 1999), and other papers of the same group, the symbols P and Q are used instead of $\mathbf{S}$ and $\mathbf{R}$. They call P the "active matrix", and Q the "passive matrix". Hurford (1989) used for the same matrices the symbols T and R, and talks about "transmission matrix" and "reception matrix"; Oliphant & Batali (1996) use the symbols s and r, and Smith (2002, 2004) uses p and r. I will use the notation and terminology introduced in chapter 2, which follows Oliphant & Batali and the standards in information theory by using the letters S and R for sender and receiver, follows Nowak et al. in using the U for the confusion matrix, and follows the mathematical convention to use boldface capitals to refer to matrices.

$$\mathbf{R} = \left(\begin{array}{c|cccc} & O_1A_1 & O_1A_2 & O_2A_1 & O_2A_2 \\ \hline w_1 & 1 & 0 & 0 & 0 \\ w_2 & 0 & 1 & 0 & 0 \\ w_3 & 0 & 0 & 1 & 0 \\ w_4 & 0 & 0 & 0 & 1 \\ N_1V_1 & 1 & 0 & 0 & 0 \\ N_1V_2 & 0 & 1 & 0 & 0 \\ N_2V_1 & 0 & 0 & 1 & 0 \\ N_2V_2 & 0 & 0 & 0 & 1 \end{array}\right),$$

where x is a single variable that describes how often the holistic strategy is used (with signals w_1, w_2, w_3, w_4) vs. how often the combinatorial strategy is used (where signals are combinations of nouns N_1, N_2 and verbs V_1, V_2). Nowak & Krakauer further assume that the confusion between holistic signals (u_h) is larger than the confusion between combinations (u_c), and that there is no confusion between the two types of strategies.

The rest of the analysis is identical to the reconstructed analysis of combinatorial phonology in chapter 4, and, with the same bit of algebra (see equations 4.4–4.6), Nowak & Krakauer (1999) can therefore draw the same conclusion: a more compositional language $L' = \{\mathbf{S}', \mathbf{R}'\}$ can always invade a population with a less compositional language $L = \{\mathbf{S}, \mathbf{R}\}$, because the fitness of L' in a population speaking L is higher than the fitness of L:

$$F(L', L') > F(L, L') > F(L, L) \;\; if \;\; (x' > x) \wedge (u_c < u_h). \tag{5.2}$$

It follows that only a fully compositional language is an Evolutionary Stable Strategy. Hence, if the assumptions implemented in this model are correct, we should expect evolution to lead to fully compositional languages. One might ask: why, then, do not all species have a compositional communication system? Nowak & Krakauer argue that not all combinations of objects and actions occur in the world or are relevant for communication and survival. If only a fraction ϕ of all combinations are relevant, the proper comparison is between a holistic system with $N = \phi nh$ words (where n is the number of objects, and h the number of actions), and a compositional system with n nouns and h verbs. Nowak & Krakauer observe that under these assumptions, compositionality is only favoured if there are more relevant events than the sum of nouns and verbs, i.e. $N \geq n + h$. They speculate that in other species compositional language was either not a possibility due to other constraints, or not favoured by selection because there were not enough relevant events to talk about (that is, N and ϕ were too small).

Given that the analyses for combinatorial phonology and compositional semantics are almost identical, it is unsurprising that the same objections apply. As I argued in chapter 4, the model considers only the advantages of the combinatorial strategies, and not the disadvantages:

- By pre-determining where in the **S** and **R** matrix the non-zero entries are, the model effectively rules out all misunderstandings except those due to acoustic noise. The model thus completely ignores the problem of coordination. If we assume that individuals speaking a holistic language do not necessarily understand the compositional signals, then compositionality cannot invade in the population (the authors do note that in such a scenario, a holistic language is an ESS as well, but do not elaborate on the consequences of this fact for their analysis).
- The model assumes there is no confusion between nouns and verbs and no confusion between the parallel holistic and compositional systems. The confusion between nouns and between verbs is only determined by how many nouns or verbs there are. Effectively, the compositional signals therefore have double the duration of holistic signals (for the reasons explored in chapter 4 the confusion between compositional signals could in fact be lower than between holistic signals in a model where both have the same duration. However, the mechanism responsible for the results in that chapter falls outwith the scope of the Nowak & Krakauer model).

However, the most important problem with the model is that compositionality here is fulfilling the exact same function as combinatorial phonology. As I argued in the previous chapter, combinatorial phonology is in fact a real solution for robustness against noise, and it can evolve through natural selection in the ways that I explored. If this process is successful, all signals will be reliably transmitted, if that is possible under the relevant constraints on noise and duration. If reliable transmission is not possible, a repertoire of signals will be close to its information-theoretic optimum (the channel capacity). In either case, compositionality, in the analysis of Nowak & Krakauer, has nothing extra to offer! That is, if signals are already composed of meaningless phonemes such that the acoustic confusion is minimised under the relevant constraints, then there is nothing more to be gained from combining meaningful words into sentences[8].

[8]The authors mention that in addition to acoustic confusion, the mistakes can be due to "incorrect assignment of meaning", which would give compositional semantics a different role than combinatorial phonology. However, in the model, the confusion probabilities of holistic, single word utterances and compositional, two word utterances are solely determined by the number of words in the repertoire, and the number of words in a sentence. This is justifiable for confusion due to acoustic noise or incorrect assignment of meaning that somehow depends on the phonological representations of words, if we assume that the word forms are distributed optimally over the available acoustic space. For incorrect assignment of meaning due to other causes, I see no reason why the confusion probabilities would be as in the Nowak & Krakauer model.

5.3.2 *Natural Selection for Learnability*

In a later paper (Nowak, Plotkin & Jansen, 2000), Martin Nowak and colleagues present a model where compositionality has a different function, which could be described as "learnability". The model views words as something akin to a disease: they spread through a population through contact between people that know the word – are "infected" – and those that do not. Nowak et al. adapt a well-known equation from mathematical epidemiology, and study the spread of words in the same holistic and verbs & nouns conditions as in the previous model. Like in epidemiology, Nowak et al.'s model also considers how the abundance of words again decreases when the individuals that know them die. This leads to an equilibrium situation where individuals know only a limited set of words, and are only partly successful in the communication with others. Because verbs and nouns can spread independently from each other, and can be combined to express previously unseen meanings, the probability of successful communication is higher in the compositional condition (if the number of relevant events is high enough, and learning nouns and verbs not much more difficult than learning holistic signals). From this analysis, the authors conclude that natural selection will favour the combinatorial strategy, once the number of relevant meanings has reached a threshold value.

There are some serious problems with this model, from the points of view of both linguistics and evolutionary biology. First of all, in the model only death limits the size of a language and the model ignores aspects like meaning, memory limitations, ambiguity and signalling errors. Such aspects are in fact more likely to limit the communicative success of a language, and existing models that take them into account give rise to equilibria with very different characteristics (e.g. Hurford, 1989; Oliphant & Batali, 1996). In these models communicative failure arises most frequently from ambiguity, rather than from the absence of the word in an individual lexicon, and compositionality will thus bring very different advantages and disadvantages. Although the extreme case of ignoring all cognitive constraints might be interesting to analyse, it is unwarranted to base an explanation of the evolution of compositionality on it, without making any reservations.

Second, the author's fail to establish the fitness advantage of an individual with a compositional language in a population *that speaks the holistic language* (the invasibility requirement from chapter 2). Instead, they calculate the average fitness in a homogeneous group. This averaging completely obscures the real problem of language evolution. The main "difficulties in imagining how language could have arisen by darwinian evolution", that the authors refer to, have to do with the problem of imagining how a syntactic individual can be successful in a non-syntactic population. The basic idea of this model, that individuals learning a combinatorial language can

generalise to unseen examples, whereas individuals learning a holistic language cannot, is correct but hardly surprising. The conceptual problems I discussed, as well as a number of technical ones[9], can perhaps be dealt with. However, as it stands, the model is completely unconvincing as an explanation of the evolution of compositional semantics.

5.3.3 *Cultural Evolution of Compositional Semantics*

A completely different class of formal models that deals with the origins of compositional semantics are based on *iterated learning*. Iterated learning models (ILM) consider what happens when the output of an induction process becomes the input for another induction process. The crucial insight is that languages will become more learnable as a consequence of their cultural transmission from generation to generation.

To see how this happens in a formal model, consider the following situation where we start with a grammar G with some set of rules $r_1 \ldots r_N$ which produces a language L with some set of sentences $s_1 \ldots s_M$.

$$G \xrightarrow[\text{production}]{} L \underset{\text{induction}}{\Longrightarrow} G' \xrightarrow[\text{production}]{} L' \underset{\text{induction}}{\Longrightarrow} G''$$

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix} \quad \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_M \end{pmatrix} \quad \begin{pmatrix} r'_1 \\ r'_2 \\ \vdots \\ r'_{N'} \end{pmatrix} \quad \begin{pmatrix} s'_1 \\ s'_2 \\ \vdots \\ s'_{M'} \end{pmatrix} \quad \begin{pmatrix} r''_1 \\ r''_2 \\ \vdots \\ r''_{N''} \end{pmatrix}$$

If we now apply an induction procedure on the language L, we induce the grammar G', which in turn produces a language L'. In a deterministic framework, we can imagine that G is a grammar that the induction algorithm cannot learn perfectly, and therefore $G \neq G'$. However, G' is likely to be a learnable grammar because it is a consequence of a learning process; when we then proceed to induce a grammar G'' from L', the induction algorithm has a good chance of finding the correct grammar, i.e. $G' = G''$. In a probabilistic framework, things are even more interesting because

[9]For instance, the comparison that the authors make between holistic and compositional strategies is inconsistent with the assumptions they present in the preceding section on lexicon dynamics. In that section they assume that word frequencies follow Zipf's law of exponential decrease with rank, whereas the "methods" section reveals – without any motivation – that for the comparison between holistic and compositional language equal frequencies are used. It is not hard to see that equal frequencies are in fact a best-case scenario for compositionality. Fitnesses depend on the probability $P(W)$ to know a relevant word W in the case of holistic language, and the probability $P(N \wedge V)$ to know the relevant noun N – verb V combination in the case of compositionality. The probabilities $P(W)$, $P(N)$ and $P(V)$ are in the equilibrium proportional to the relative frequencies of the words, nouns and verbs, such that the average $P(W_e)$ over all relevant events e is independent from the frequency distribution, but the average $P(N_e \wedge V_e)$ has its optimum at an equal distribution. Consequently, the fitness advantage of compositionality $\frac{1}{E}\sum_e^E (P(N_e \wedge V_e) - P(W_e))$ has its optimum at equal frequencies.

here the accuracy of learning will increase over the course of a number of generations (Briscoe, 2000a, 2002b; Kirby, 2001; Zuidema, 2003a), as will be explored in more detail in chapter 6.

The relevance of this for the evolution of compositionality becomes clear if one considers that – as in the model discussed in the previous section – compositionality aids learnability. That is, the long, idiosyncratic lists of signals of holistic languages are difficult to learn, because the learner needs to see an example of every single instance. In contrast, compositional languages allow one to *generalise* from a fair number of training samples, to (possibly infinitely) many more unseen cases. In cultural transmission with a bottle-neck (a "poverty of the stimulus"), holistic language will therefore be unstable, and the language will therefore continue to change until it has become learnable and hence compositional. Batali (1998) and Kirby (2000) were the first to demonstrate in formal models that cultural transmission, with learners learning from learners, can therefore yield a compositional semantics.

I believe the iterated learning models make an important point, and will explore in the next chapters the important implications of these models for debates in linguistics on innateness, learnability and language universals. One weakness of the models studied so-far, is that the relations with formal models and results in theoretical linguistics and learnability theory have not been sufficiently explored. More importantly, however, the problem with the models as explanations for the evolutionary transition to compositional semantics, is that they – in two different ways, and to different degrees – already presuppose the existence of what they are meant to explain.

First, In Kirby's original models (Kirby, 2000, 2001, 2002a), the agents in the simulation come equipped with the representational capacity for context-free grammars (enriched with semantics), and a specialised learning algorithm to induce such grammars from example sentences. An important question is whether it is necessary and reasonable to assume that such learning abilities existed in early hominids before the object of learning, a compositional language, existed. It only is, of course, if one could demonstrate that the ability to induce context-free grammars, or any other sufficiently expressive formalism, is not language-specific, but part of the general learning abilities that prelinguistic hominids can be reasonably assumed to have had. Iterated Learning Models that use different formalisms and learning rules (e.g. Batali, 1998; Smith, 2002; Brighton, 2002) seem less biased toward compositionality. Unfortunately, it is hard to judge what kind of learning biases are reasonable.

Second, the success of an ILM is dependent on the probability that some kind of structure, that the learning algorithm can detect, arises by chance. In Kirby's (2000) model, random strings of characters are generated for (compound) meanings that cannot be expressed by the current

grammar. E.g. if the meaning *loves*(*Mary*, *Tünde*) cannot be expressed using the rules of the current grammar, a random string *abacdddbe* might be generated to express it ("invention"). For a related meaning, e.g. *hates*(*Mary*, *Tünde*) another random string will be generated. Only when by chance both strings share a substring, will the learner induce the corresponding compositional rules[10]. In the simulations, the probability that this happens is relatively high, because of specific choices for the number of possible characters and the random string length. It is unclear, however, how realistic these choices are. Again, it is difficult to judge what kind of invention biases are reasonable.

Kirby and colleagues are aware of these limitations. Kirby (2000) emphasises that the model shows, contra Pinker & Bloom (1990), that there are processes other than Natural Selection capable of explaining complex patterns in natural language. I believe it serves well as such a counterexample, and that the mechanism at work in the iterated learning model will play an important role in understanding the features of natural languages. However, for similar reasons as explored in chapter 4, I think it is worthwhile to explore the fitness effects of increased compositionality, if only to evaluate whether natural selection will work with such alternative processes, or against it.

5.4 Model Description

The model of this chapter follows Nowak & Krakauer (1999) in focusing on the interaction between noise robustness and compositionality, but rather than viewing compositionality as a *strategy* to impose noise robustness, it views compositionality as a *side-effect* of optimising signals for robustness. An assumption in the model – and a crucial difference with Nowak & Krakauer's – is that not all mistakes are equally bad. If an interpretation is wrong but close, it is worse than the correct interpretation, but better than a completely different interpretation. Another crucial difference with Nowak & Krakauer's, is that in my model agents always only know a single language.

5.4.1 Hill-climbing

Similarities between meanings are reflected in a value matrix $\mathbf{V}$, that describes the value of each interpretation for any given intention, as I discussed in chapter 3. The expected payoff between a speaker i and a hearer j then becomes (equation 3.1):

$$w_{ij} = \mathbf{V} \cdot \left(\mathbf{S}^i \times \left(\mathbf{U} \times \mathbf{R}^j\right)\right) \tag{5.3}$$

[10]Kirby makes no claim that the learning algorithm used models human language acquisition. Kirby studies how the language changes over the course of many generations; he has deliberately simplified the learning algorithm, inspired on Stolcke (1994), to reduce the computational and conceptual complexity of the model.

In this formula, "$\times$" represents the usual matrix multiplication and "$\cdot$" represents dot-multiplication (the sum of all multiplications of corresponding elements in both matrices; the result of dot-multiplication is not a matrix, but a scalar). There are M different meanings that an individual might want to express, and F different signals (forms) that it can use for this task. $\mathbf{S}$ is a $M \times F$ matrix that gives for every meaning m and every signal f, the probability that the individual chooses f to convey m. Conversely, $\mathbf{R}$ is a $F \times M$ matrix that gives for every signal f and meaning m, the probability that f will be interpreted as m. $\mathbf{U}$ (of dimension $F \times F$) gives for every uttered signal f the probability it is perceived as f', $\mathbf{V}$ (of dimension $M \times M$) gives for every intention m the payoff of an interpretation m'.

Based on this measure, I use some simple hill-climbing heuristics to improve the communication. Specifically, I will report simulation results with three types of hill-climbing:

Global optimisation of a probabilistic lexicon: in this condition, there is a single $\mathbf{S}$ and a single $\mathbf{R}$ matrix. The matrices are initialised with random, real-valued entries between 0 and 1 (and rows or columns normalised). At every step in the simulation, an entry in one of the matrices is chosen at random, a small degree of noise is added (from a Gaussian $\mathcal{N}(0, \rho)$, i.e. with mean $\mu = 0$ and standard deviation $\sigma = \rho$, the "learning rate" parameter) and the expected payoff $w = \mathbf{V} \cdot (\mathbf{S} \times (\mathbf{U} \times \mathbf{R}))$ is measured. If w is at least as large as before, the change is kept, otherwise it is reversed. The matrices describe the average production and interpretation probabilities in a population; this condition thus corresponds to the unrealistic scenario where communication is optimised for the benefit of the whole population. These simulations mirror the analytic calculation of maximum fitness in Nowak & Krakauer (1999); the main difference is the $\mathbf{V}$-matrix.

Local optimisation of a probabilistic lexicon: in this condition, a population (of size N) of individuals is modelled with everyone having her own $\mathbf{S}$ and $\mathbf{R}$ matrices. The matrices are initialised with random, real-valued entries between 0 and 1 (and rows or columns normalised). At every step in the simulation, a random speaker i and a random hearer j are selected, and the expected payoff between them is calculated (equation 5.3). As before, an entry in either the speaker's $\mathbf{S}$ or the hearer's $\mathbf{R}$ is chosen at random, a small degree of noise is added, and the change is kept if $w'_{ij} > w_{ij}$. This version of my model is very similar to simulations reported in Nowak & Krakauer (1999); the main difference is again the $\mathbf{V}$-matrix.

Local optimisation of a deterministic lexicon: this condition is identical, except that the values of the entries are restricted to either 1 or 0. That means that they are deterministic encoders

and decoders, which can be shown to always perform better than their stochastic versions (Shannon, 1948; Plotkin & Nowak, 2000). This simplification allows for an enormous speed-up of the simulation (using the algorithms in appendix B of this thesis), allowing for many more and larger-scale simulations. The random change in the hill-climbing procedure is to move the only 1 in a random row or column to a random other position in that row or column.

The motivation the local optimisation conditions is (i) that they are fast and straightforward to implement; (ii) that they work well, and give, if not the optimum, a good insight in the characteristics of the optimal communication system; and (iii) that they consider the invasibility of linguistic traits in a population where they are rare, and thus show possible *routes* to (near-) optimal communication systems, and in a sense form an abstraction for both learning and evolution[11]. If the simulations converge such that all individuals have the same $\mathbf{S}$ and the same $\mathbf{R}$ matrix, these particular matrices define an Evolutionary Stable Strategy (with respect to a strategy set containing all strategies that are 1-step mutations from the ESS).

5.4.2 *Semantic Similarity and Acoustic Confusability*

The $\mathbf{V}$ and $\mathbf{U}$ matrices can be chosen to reflect all kinds of assumptions about the signal and meaning space. A $\mathbf{V}$ could theoretically be constructed from empirical observations, if one could list the "meanings" available for communication, and measure the payoff from all of these meanings as interpretations in the context of all these meanings as intentions. Alternatively, one could estimate these values from a measure of *semantic similarity* and a function relating similarity to payoff. Similarly, the $\mathbf{U}$ matrix could be constructed from empirical observations, if one could list all the possible "signals" available for communication, and measure the probability of confusion between each possible produced signal and each possible perceived signal[12].

In this chapter I will not be concerned with measures of semantic similarity or confusability. Instead, I will study the consequences of some specific choices for $\mathbf{V}$ and $\mathbf{U}$ that reflect qualitatively different assumptions on (i) whether all meanings are equally valuable or not, and (ii)

[11] There are no deep philosophical reasons for the difference in the way I deal with invasibility in this chapter and the previous. The difference is an arbitrary modelling choice. In this chapter, an explicit population of agents is modelled, reflecting a preference for such explicit models in the Artificial Life community (although the abstractions that I do make – for instance, using hill-climbing rather than an observational or reinforcement learning paradigm – are quite unlike much Artificial Life work); in chapter 4, more in line with work in evolutionary game theory, invasibility is dealt with without simulating a population.

[12] For simplicity, I assume throughout this thesis that the set of meanings available as "intentions" is the same as the set available as "interpretations", and that the set of signals available as "articulations" is the same as the set available as "acoustic perceptions". This is in contrast to work on signalling games in the tradition of Lewis (1969), where the intentions ("situations") are from a different set than the interpretations ("actions"). For empirical estimates of the $\mathbf{V}$, $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{R}$ matrices it might be more convenient to distinguish between all these cases, but the essential apparatus developed in this chapter will still be available.

whether or not there is a topology in the meaning and signal space, and if so, of which dimensionalities.

For the $\mathbf{V}$ matrix, I will look at the following conditions. First, for the diagonal elements (the correct interpretations) the **homogeneous** condition assumes that there are no qualitative differences between meanings. Hence, all diagonal values are 1. In the **heterogeneous** condition, in contrast, some meanings are more important than others. In the simulations under this condition, I assign a random value v to each meaning, which defines the corresponding diagonal element in $\mathbf{V}$. For the off-diagonal elements in the $\mathbf{V}$ matrix, I consider three conditions:

0d: There is no topology in the meaning space. Hence, every wrong interpretation is equally bad: all off-diagonal values in $\mathbf{V}$ are 0. A *homogeneous, 0d* $\mathbf{V}$ matrix is a $M \times M$ unit matrix as in figure 5.2(a).

1d: There is a 1-dimensional topology in the meaning space. The position of a meaning in that space corresponds to its index in the matrix. That is, I assume that if m_3 is the intended meaning, interpretation m_3 gives the highest payoff, m_2 and m_4 a lower but non-zero payoff, m_1 and m_5 an even lower payoff and so forth. An example of a *homogeneous, 1d* $\mathbf{V}$ matrix is given in figure 5.2(b). An example of a *heterogeneous, 1d* $\mathbf{V}$ matrix is given in figure 5.2(c). The exact values of the entries in $\mathbf{V}$ are defined below.

2d: There is a 2-dimensional topology in the meaning space. Here I assume the meaning-space is a perfect square. Each of the positions in this space is labeled with a unique number, as illustrated in figure 5.3(a), which is the index of each particular meaning in the $\mathbf{S}$, $\mathbf{R}$ and $\mathbf{V}$-matrices. Figure 5.3(c) is a $\mathbf{V}$-matrix that reflects such a 2d topology.

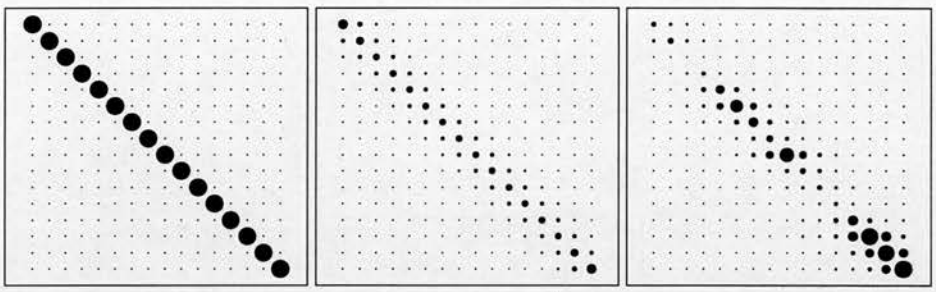
Similarly, I will look at $\mathbf{U}$ matrices that reflect no topology, a 1-dimensional or a 2-dimensional topology in the signal space. An example of the labeling of signals in a 2-dimensional signal space, with the indices that are used in the $\mathbf{S}$, $\mathbf{R}$ and $\mathbf{U}$ matrices, is given in figure 5.3b. In this chapter I will not look at heterogeneous $\mathbf{U}$ matrices, but the generalisation is easily made, and in chapter 7 I will look at one simple example of a simulation under this condition.

The precise values of entries of the $\mathbf{V}$ and $\mathbf{U}$ matrices are given by the following equations:

$$\mathbf{V}(p,q) = v/(1+d(p,q)), \tag{5.4}$$

$$\mathbf{U}(p,q) = 1/(1+d(p,q)), \tag{5.5}$$

where $v = 1.0$ in the homogeneous $\mathbf{V}$ condition, and a random value ($0.0 < v \leq 1.0$) in the "heterogeneous" $\mathbf{V}$ condition; without a topology ("0d"), the off-diagonal elements in $\mathbf{U}$ or $\mathbf{V}$



(a) $\mathbf{V}$: $0d$ (b) $\mathbf{V}$: $1d$ homogeneous (c) $\mathbf{V}$: $1d$ heterogeneous

Figure 5.2: Examples of $\mathbf{V}$ matrices that reflect different qualitative assumptions on the structure of the meaning space. On the vertical axis are all meanings as intentions, on the horizontal axis all meanings as interpretations. In (a) all meanings are equally valuable and interpretations only give a payoff if they are absolutely correct. In (b) all meanings are equally valuable and correct interpretations give the highest payoff (the diagonal entries), but slightly wrong interpretations still give some payoff. In (c) different meanings have different values, and slightly wrong interpretations still give part of those payoffs. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are plotted as a small dot.

are 0 (i.e. $d(p,q) = \infty$ if $p \neq q$). With a topology, $d(p,q)$ gives the squared Euclidean distance between the positions of the two meanings or signals p and q. After these values are set, the rows of both $\mathbf{U}$ and $\mathbf{V}$ matrices are normalised[13] such that the values of each row add up to one.

In the 1-dimensional condition the position of a meaning or signal is simply defined as its index. In the 2d condition, the meaning and signal spaces are 2-dimensional surfaces of size $(\sqrt{M} \times \sqrt{M})$ or $(\sqrt{F} \times \sqrt{F})$ (see figure 5.3a,b). Each of the positions in those spaces is labeled with an index, with 0 in the bottom left corner, 1 one position higher and so-forth. When given an index, we can calculate the corresponding x- and y-coordinates as follows:

- The x-coordinate is given by the largest integer smaller than the root of the index: $x = int(\sqrt{i})$;
- The y-coordinate by: $y = i$ *modulo* x.

5.4.3 *Performance Measures*

I monitor the behaviour of the model with two measures. The first is the average payoff, as given by equation (5.3), averaged over all individuals interacting with all other individuals, both as speaker and as hearer. The second is a measure for the degree of topology preservation between

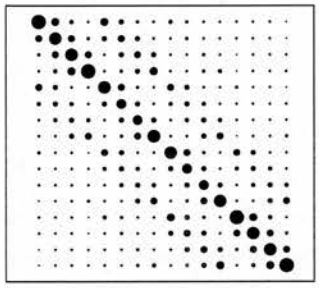
[13]The normalisation of the $\mathbf{V}$ matrix is in fact unnecessary, and introduces an unfortunate boundary effect: meanings with fewer neighbours have slightly higher diagonal values, as is apparent in the 1-dimensional condition at the top-left and bottom-right corners (figure 5.2b) and in the 2-dimensional condition in the slightly different values for meanings with 2, 3 or 4 neighbours (figure 5.3c). The effect is very small, however, and the simulations were therefore not redone.

3	7	11	15
2	6	10	14
1	5	9	13
0	4	8	12

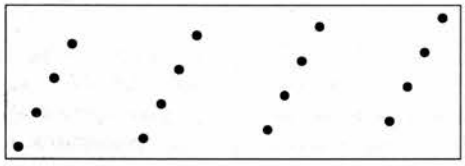
(a) The 4×4 meaning space, with the indices of 16 meanings

6	13	20	27	34	41	48
5	12	19	26	33	40	47
4	11	18	25	32	39	46
3	10	17	24	31	38	45
2	9	16	23	30	37	44
1	8	15	22	29	36	43
0	7	14	21	28	35	42

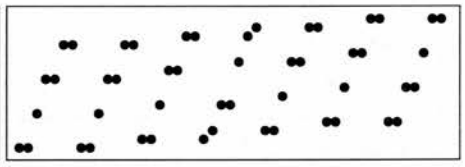
(b) The 7×7 signal space, with the indices of 49 signals



(c) The 16×16 $\mathbf{V}$-matrix, giving for each of 16 possible intentions, the payoff of each of 16 possible interpretations, and reflecting the $2d$ topology of the meaning space. ($\mathbf{V}$:2d homogeneous)



(d) A 16×49 $\mathbf{S}$-matrix, with 16 meanings on the vertical axis and 49 signals on the horizontal axis. The bottom left point represents meaning 0 and signal 0.



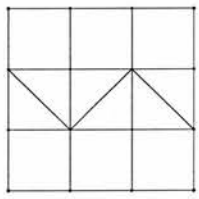
(e) A 16×49 $\mathbf{R}^T$-matrix (transposed for easy comparison), with 16 meanings on the vertical axis and 49 signals on the horizontal axis.

42	44	46	48
28	30	32	34
14	16	18	20
0	2	4	6

(f) The 4×4 meaning space, with the indices of 16 corresponding signals, according to $\mathbf{S}$

12	12	**13**	14	**14**	15	**15**
12	12	13	13	14	15	15
8	8	**9**	10	**10**	11	**11**
8	8	9	5	10	11	7
4	4	**5**	5	**6**	7	**7**
0	0	1	2	2	3	3
0	0	**1**	1	**2**	3	**3**

(g) The 7×7 signal space, with the indices of 16 corresponding meanings, according to $\mathbf{R}$ (in boldface the preferred signals in the $\mathbf{S}$ matrix).



(h) The 4×4 meaning space, with neighbours in signal space, according to $\mathbf{R}$, connected by edges.

Figure 5.3: 2d meaning and signal spaces and visualising topology preservation.

the meaning space and the signal space in the emerging languages. Following Brighton (2002), I use the correlation ("Pearson's r") between the distance between each pair of meanings and the distance between the corresponding signals:

$$r = \underset{m,m' \in M}{\text{correlation}} \left(D\left(m, m'\right), D\left(S[m], S[m']\right)\right), \tag{5.6}$$

where $S[m]$ gives the most likely signal used to express m according to $\mathbf{S}$. This measure gives a value 1 when for every meaning–signal pair the coordinates in meaning space and signal space are identical (or equivalent under mirroring and rotation), and 0 when the mapping is random (Brighton, 2003).

For 2-dimensional meaning spaces I also visualise the topology preservation by plotting all meanings as nodes in a meaning space, and connecting those nodes where the corresponding signals are neighbours (one of maximum four) in signal-space. Figure 5.3 illustrates this visualisation technique, for the given $\mathbf{S}$ and $\mathbf{R}$-matrices (figure 5.3d,e; the origin of these matrices is not relevant here, but they result from the same simulation as reported in figure 5.12a). The $\mathbf{S}$ matrix associates meanings with signals; in figure 5.3f the indices of the signals are plotted at the locations of the corresponding meanings in meaning space. The $\mathbf{R}$ matrix associates signals with meanings; in figure 5.3g the indices of the meanings are plotted at the locations of the corresponding signals in signal space.

Finally, in figure 5.3h the representation I will use in this chapter is given. Here, points in meaning space are connected if the corresponding signals are neighbours in signal space. Thus, signals 5, 6, 12 and 13 (the 4 signals in the top left corner, see 5.3b), are all interpreted as meaning 12 (see the top left corner in 5.3g). The neighbours of these signals are 4, 11, 18, 19 and 20, which are interpreted as meanings 8, 8, 9, 13 and 13 respectively. Therefore, in the final representation meaning 12 is connected with meanings 8, 9 and 13. Using this representation, the topology preservation between meaning space and signal space, which is almost perfect in this example, is immediately clear.

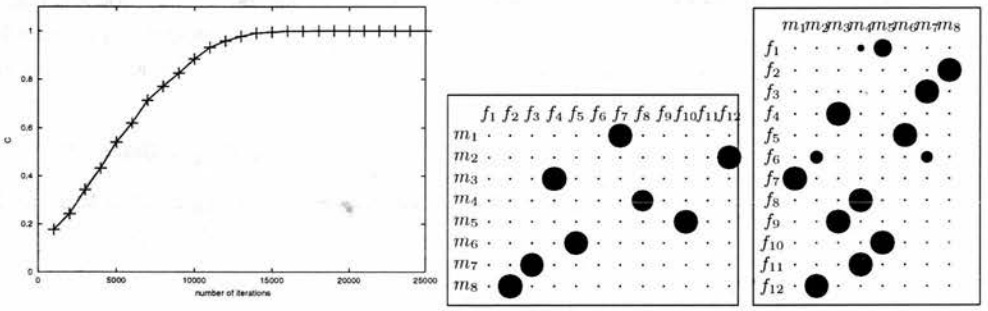
5.5 Properties of the Optimal Lexicon

To give an idea of the properties of the optimal lexicon – that might or might not be an evolutionary stable strategy – I will in this section present results from some very simple simulations with the frequency-independent hill-climbing heuristic (the "global optimisation of a probabilistic lexicon" condition).

Throughout this chapter, I assume that there are more signals available than meanings to express. This is a fairly strong assumption, but there are two good reasons for choosing it as a starting point. First, if signals are from a continuous space (as in chapter 4), and meanings from a finite (non-recursive), discrete space (such as first-order predicate logic), then there will be (infinitely) many more possible signals than meanings. Although the number of *possible* signals in the simulation is finite, the number of *usable* signals will – as in reality – be determined by the amount of noise. Second, the assumption simplifies the dynamics considerably and, if the noise level is sufficiently low, guarantees the existence of at least one Evolutionary Stable Strategy in the "local optimisation" conditions. The dynamics of models with different assumptions on the meaning- and signal-spaces (including continuous and hierarchically structured spaces) remain to be explored (a start has been made with a related model by Matina Donaldson, p.c.).

5.5.1 Categorical Meanings, Noise-free Signalling

The simplest case is where there is categorical, noise-free communication. That is, every meaning is unique and has no relation with other meanings, and signals are perceived as they are uttered. These conditions ($\mathbf{V} : 0d$, $\mathbf{U} : 0d$) are described with a $\mathbf{V}$ and $\mathbf{U}$ that are both unit matrices (matrices with 1's on the diagonal, and 0's everywhere else).



(a) Development of communicative success over 25000 iterations

(b) S matrix of a random individual, showing for each meaning (vertical) the probability that she will use any of the signals (horizontal) to express it.

(c) R matrix of the same individual, showing for each signal (vertical) the probability that she will choose any of the meanings (horizontal) as its interpretation.

Figure 5.4: The optimised lexicon in a population under categorical, noise-free conditions. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are plotted as a small dot. ($\mathbf{V} : 0d$, $\mathbf{U} : 0d$, $M = 8$, $F = 12$, $N = 3$, $n = 0.1$).

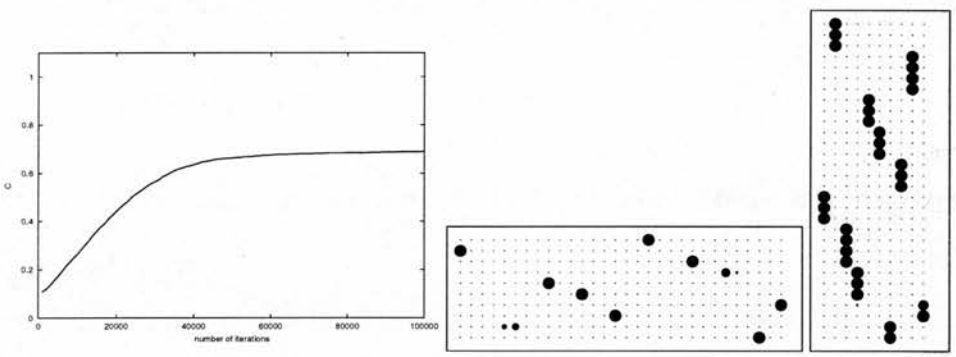
Optimising a population's lexicon under these conditions using the hill-climbing algorithm described above, gives results as in figure 5.4. The average payoff increases steadily and reaches

the optimal value (1.0). The **S** matrices in the population have maximal probability (= 1.0) for a specific signal (horizontal) for each of the meanings (vertical), and probability 0 for all other signals. In the **R** matrix these signals (vertical) are interpreted as the "correct" meanings. Because there are more possible signals than meanings, some signals (f_1, f_6, f_9, f_{11}) are never used and have arbitrary interpretations.

This simple simulation illustrates two properties of the optimal lexicon: *specificity*, "one unique signal for every intention, and one unique interpretation for every used signal", if $M \leq F$, and *coherence*, "everyone in a population uses the same signal for the same meaning".

5.5.2 *Categorical Meanings, Noisy Signalling*

If there is noise on the signal (due to a noisy environment and sensory limitations of the hearer), the hearer will sometimes hear a different signal than the speaker uttered. We can model this by introducing non-zero, off-diagonal entries in the matrix **U**. Here, I consider only the simplest case, where signals vary on one axis determined by their index (**U** : 1d).



(a) Development of communicative success over 100000 iterations

(b) **S** matrix of a random individual.

(c) **R** matrix of the same individual.

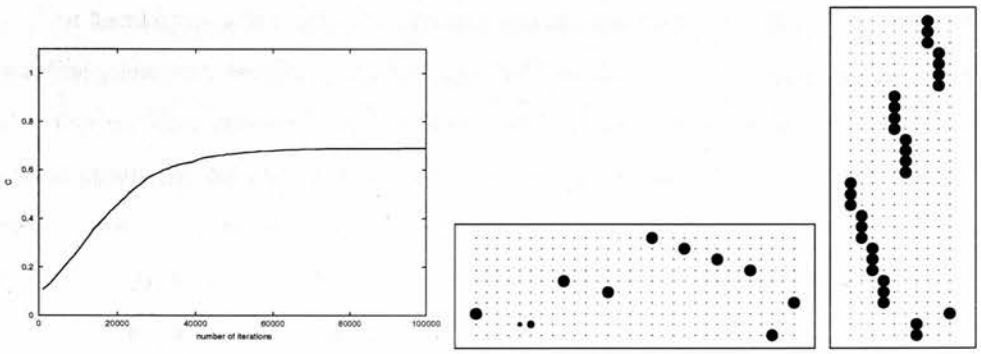
Figure 5.5: A local optimum of the lexicon in a population under categorical, noisy conditions (**V** : 0d, **U** : 1d, $M = 10$, $F = 30$, $N = 3$, $n = 0.1$).

Under these conditions, we expect a lower average payoff and **S** and **R** matrices that somehow minimise the chance of confusion. Figure 5.5 shows that this is indeed what happens. The **S** matrix shows that for every meaning, there is a prototype signal that individuals use. For these prototype signals and their direct neighbours, the interpretation is the "correct" meaning. Little clusters of neighbouring signals are all interpreted in the same way, such that prototype signals are maximally distinct from each other. Thus, in addition to specificity and coherence, *distinctiveness*

is a property of the optimal lexicon when the signalling is noisy. Note that, even though there are many more signals than meanings, all signals have a specific "best" interpretation.

5.5.3 *Semantic Similarities & Noisy Signalling*

If we include in the model the assumption that not only signals have similarity relations, but also meanings relate to each other, we can identify a fourth criterion of the optimal lexicon: *regularity*. Figure 5.6 shows results that are obtained by running the same hill-climbing algorithm, with $\mathbf{V} : 1d$ and $\mathbf{U} : 1d$.



(a) Development of communicative success over 100000 iterations

(b) S matrix of a random individual

(c) R matrix of the same individual

Figure 5.6: Local optima for S and R under semantic similarities, noisy signalling conditions ($\mathbf{V} : 1d$, $\mathbf{U} : 1d$, $M = 10$, $F = 30$, $N = 3$, $n = 0.1$).

The local optima found by the hill-climbing algorithm show not only specificity, coherence and distinctiveness, but also *partial regularity*: "similar signals tend to have similar meanings", such that misinterpretations are still better than a random interpretation. The solution found is a local optimum; the observed patterns suggest that the globally optimal lexicon is maximally regular: with the parameters of the simulations in figure 5.6, meaning m_1 would be expressed with signal f_1, and signals f_2 to f_3 are interpreted as m_1; m_2 is expressed with f_5, and f_4 to f_6 are interpreted as m_2 etc. This optimum is not found in this simulation; however, in the local optimum of figure 5.6 neighbouring clusters of signals are, with only a few exceptions, associated with neighbouring meanings. Measuring the degree of regularity r shows that it is consistently higher under conditions with semantic similarities than without.

5.5.4 *Properties of the Optimal Lexicon*

These simulations illustrate that the optimal lexicon must have the following properties (provided that $M << F$, and that the off-diagonal U and V values are relatively low):

specificity: every meaning has exactly one signal to express it and vice versa (i.e. no homonyms, and no real synonyms: if different signals have the same meaning they are very similar to each other). In the representation I use in this chapter, this property is present if there is only a single full-sized circle in each row of the $\mathbf{S}$ and $\mathbf{R}$-matrices (or each column of the $\mathbf{R}^T$-matrix).

coherence: all agents agree on which signals to use for which meanings, and vice versa. This property is present if all $\mathbf{S}$ and $\mathbf{R}$-matrices in a population are identical.

distinctiveness: the used signals are maximally dissimilar to each other, so that they can be easily distinguished. In the $\mathbf{U} : 1d$ condition, this property is visible in the $\mathbf{S}$-matrix if the circles are maximally dispersed over the width of the matrix, and in the $\mathbf{R}$-matrix as little clusters of signals that all have the same interpretation. In the $\mathbf{U} : 2d$ condition, this property is visible in the distribution of preferred signals from the $\mathbf{S}$-matrix in signal space (see, for instance, the boldface meaning-signal pairs in figure 5.3g).

regularity: in the mapping between meanings and signals there is a *preservation of topology*, i.e., similar signals tend to have similar meanings. In the $(\mathbf{V} : 1d, \mathbf{U} : 1d)$ condition, this property can be seen in the $\mathbf{S}$ and $\mathbf{R}$-matrices as local staircase-like patterns. In the $\mathbf{V} : 2d$ condition, it can be visualised using the technique I described in section 5.4.3. In all cases, it can be roughly measured with the r correlation measure of equation (5.6), where a value of $r \approx 1$ indicates perfect topology preservation.

These properties are quite different from the properties of the lexicon of any natural language, which of course has not been globally optimised, where the $\mathbf{V}$, $\mathbf{U}$, M and F are all quite different, and where the interpretation of each word is crucially dependent on the context. However, these properties do follow naturally from the assumptions about topology in the meaning and signal spaces I made in this model. If one were to design a code for communication over a noisy channel, without context, the same properties would emerge: (i) senders should consistently use a unique signal for each meaning they need to express, and receivers should decode each signal with a unique meaning; (ii) everyone in the population should agree on the same code; (iii) when there is a range of signals available, the signals used in the code should be well-spread over the available space, and signals received with some distortion should be decoded as the most likely (nearest) prototype; (iv) if there is freedom in how to organise the mapping, the damage of unavoidable confusion should be minimised, such that misinterpreted signals receive the next best interpretation.

The exact shape of the optimal lexicon will depend on the specific choices of $\mathbf{V}$, $\mathbf{U}$, M and F. For the purposes of this chapter, a qualitative understanding of its properties are sufficient. The real issue for this chapter – and a recurrent theme throughout this thesis – is how traits that might or might not be beneficial for the group, can *invade* in a population, i.e. emerge through *local* optimisation. In the next sections I will study a simulation of a population of agents, where each agent tries to optimise her success in communicating with a randomly picked other agent.

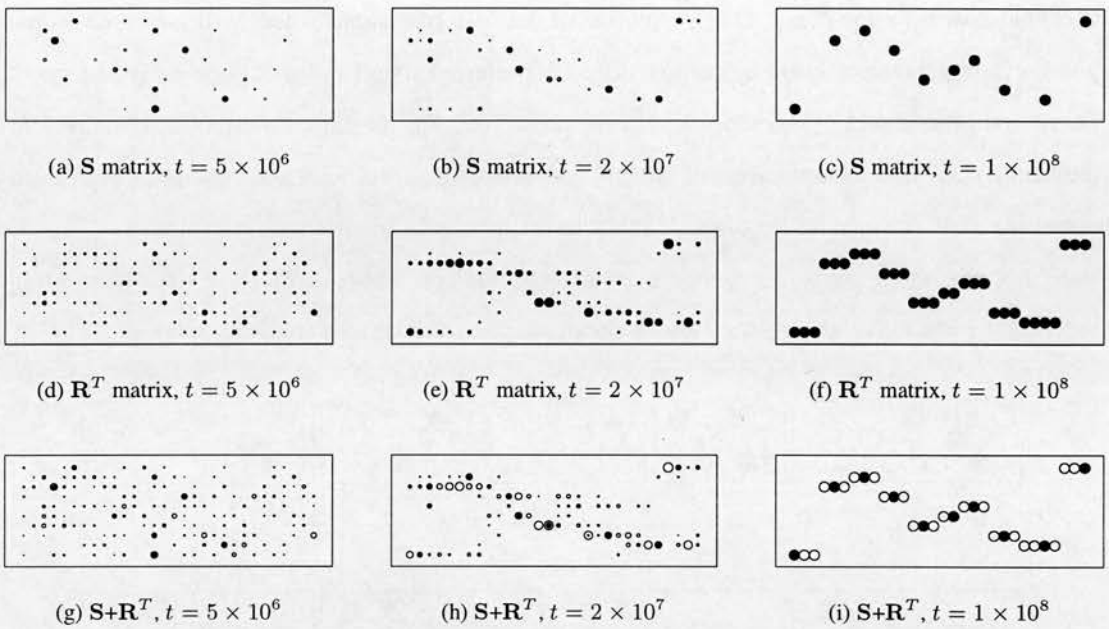
5.6 Local Optimisation of a Probabilistic Lexicon

Figure 5.7 shows results from a simulation with the same parameters as in figure 5.6, but with local hill-climbing in a population with $N = 40$ agents and a higher learning rate (the random change in the hill-climbing algorithm is from a Gaussian distribution with mean 0 and standard deviation $\rho = 1.0$). The figure shows $\mathbf{S}$ and $\mathbf{R}$ matrices from one random individual at three points in the simulation: after 5×10^6 and 2×10^7 iterations, and in the stable equilibrium configuration (after almost 1×10^8 iterations)[14].

The lexicon that develops shows all 4 characteristics of the optimal lexicon. In the $\mathbf{S}$ matrix at equilibrium (from around $t = 1 \times 10^8$) every meaning is always expressed by one unique signal; in the $\mathbf{R}$ matrix, that signal is always interpreted with the correct meaning (**specificity**). At equilibrium, all agents have the same $\mathbf{S}$ and $\mathbf{R}$ matrices (**coherence**). It is not difficult to see why these properties emerge in a local optimisation set-up. Consider that at any point in the simulation, there is for each meaning one specific signal that gives the highest chance of being understood (this is because all values are continuous, and the values in the $\mathbf{R}$ matrices are therefore never exactly equal). For every speaker, it therefore pays off – on average – to use that specific signal in the $\mathbf{S}$ matrix. Conversely, for every signal there is one specific meaning that is the most probable interpretation. For every hearer, it therefore pays off – on average – to use that specific meaning in the $\mathbf{R}$ matrix. Only a lexicon with specificity and coherence can therefore be an *Evolutionary Stable State.*

In the $\mathbf{S}$ matrix at equilibrium, the preferred signals are (almost) maximally dispersed. Figure 5.7i, which shows the $\mathbf{R}^T$ matrix from figure 5.7f (with open circles) overlain with the $\mathbf{S}$ matrix of figure 5.7c (with closed circles), shows that in the $\mathbf{R}$ matrix, each of these preferred signals (except at the edges) is at the centre of a little cluster of signals that are all interpreted with

[14] An interesting question is whether 10^8 iterations is too long for these results to be relevant for scenarios of language evolution. That is, whether or not there has been *sufficient time* (criterion 6 in chapter 2) for the mechanism modelled here to have played a role in shaping natural language. This is an important issue, but one that goes beyond the scope of this thesis. Answering this question requires first of all more robust results on how the time to convergence depends on the many parameters in the model, and second, a more concrete interpretation of what the optimisation steps correspond to in the real world (genetic mutations, selection, learning events).



(a) $\mathbf{S}$ matrix, $t = 5 \times 10^6$ (b) $\mathbf{S}$ matrix, $t = 2 \times 10^7$ (c) $\mathbf{S}$ matrix, $t = 1 \times 10^8$

(d) $\mathbf{R}^T$ matrix, $t = 5 \times 10^6$ (e) $\mathbf{R}^T$ matrix, $t = 2 \times 10^7$ (f) $\mathbf{R}^T$ matrix, $t = 1 \times 10^8$

(g) $\mathbf{S}$+$\mathbf{R}^T$, $t = 5 \times 10^6$ (h) $\mathbf{S}$+$\mathbf{R}^T$, $t = 2 \times 10^7$ (i) $\mathbf{S}$+$\mathbf{R}^T$, $t = 1 \times 10^8$

Figure 5.7: Development of specificity, coherence, distinctiveness and regularity in the lexicon of a population under semantic similarities, noisy signalling conditions. At each time-step a random speaker interacts with a random hearer and one of them performs a single hill-climbing step to improve the communication. In this graph, $\mathbf{R}$ matrices are transposed, such that in both $\mathbf{S}$ and $\mathbf{R}^\mathbf{T}$ matrices meanings are on the vertical axis and signals on the horizontal axis. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are not plotted. (a-c) show the $\mathbf{S}$ matrices, (d-f) the $\mathbf{R}^T$ matrices, and (g-i) the $\mathbf{R}^T$ matrices (with open circles) with the $\mathbf{S}$ matrices (with closed circles) overlain. Parameters: $\mathbf{V} : 1d$, $\mathbf{U} : 1d$, $M = 10$, $F = 30$, $N = 40$, $\rho = 1.0$.

the same meaning (**distinctiveness**). The reason this property emerges is slightly more subtle. One way to see why it is inevitable is as follows. Consider a population where the *specificity* and *coherence* of a lexicon have been established, but *distinctiveness* has not. Assume futher that there are many more signals than meanings, such that there must be signals that are not the preferred signal for any meaning (the *non-preferred signals*). Moreover, because of the noise on transmission (as modelled by the $\mathbf{U}$ matrix), a signal s as the preferred signal for meaning m is not always perceived correctly, but sometimes perceived as a neighbouring signal s'.

In this situation (a concrete example is given in figure 5.8a), it pays off for an agent to shift the interpretation of any non-preferred signal s' in the $\mathbf{R}$ matrix, to the same meaning $R[s] = m$ as that of a neighbouring preferred signal s. The same logic applies to non-preferred signals that are further away than distance 1 from a preferred signal. Consequently, at least one little cluster in the $\mathbf{R}$ matrix forms, as in figure 5.8c. Given that situation, it now pays off for an agent to shift the preferred signal in the $\mathbf{S}$ matrix to the center of the cluster, or, if the cluster is at the edge

of signal space, to this edge. This movement of the preferred signal, then, will mean that some non-preferred signals are now closer to a different preferred signal and should be moved in the $\mathbf{R}$ matrix, and the same process repeats. This sequence of steps does not describe the dynamics in the simulation, where distinctiveness already appears before specificity and coherence have been established. It does, however, show that a non-distinctive lexicon is not an evolutionary stable state, because there exists a sequence of changes to lexicon that each improve the fitness of an individual and can therefore *invade* in the population, as is illustrated in figure 5.8.

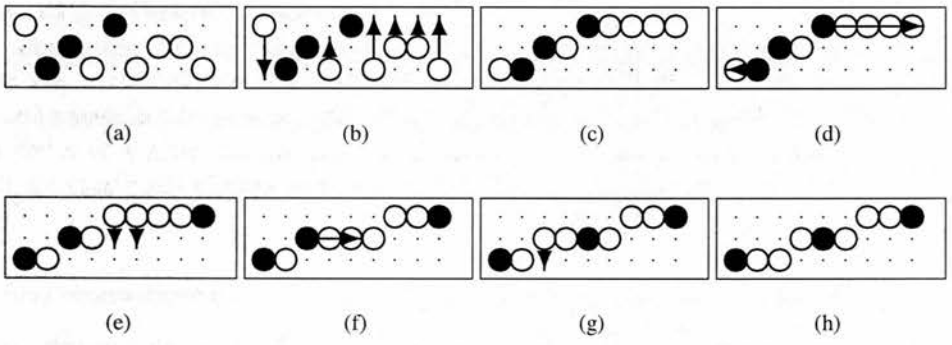


Figure 5.8: Only distinctive lexicons are evolutionary stable states. Consider a coherent, specific lexicon as sketched in (a). This diagram shows the same $\mathbf{S}+\mathbf{R}^T$-representation as in figure 5.7(g-i), with closed circles representing the preferred signals in the $\mathbf{S}$-matrix, and both open and closed circles representing the interpretation for each signal in the $\mathbf{R}$ matrix. This lexicon does not define an evolutionary stable state, because all the changes to non-preferred signals in the $\mathbf{R}$-matrix, as indicated with vertical arrows in (b), are beneficial for an individual, even if the whole population has adopted (a). These changes can therefore invade in the population. If the whole population would adopt each of these change, (c) would describe the new population lexicon. This is not a stable state either, because changing the preferred signals in the $\mathbf{S}$-matrix, as indicated with horizontal arrows in (d) again benefits an individual even if the whole population uses lexicon (c). With a series of similar changes that can all invade the population as sketched in (e-g), we end up with the maximally distinctive lexicon in (h). This lexicon does define an evolutionary stable state.

In final observation in the matrices of figure 5.7, is that, with 3 exceptions, all signal-clusters have neighbouring signal-clusters that express a neighbouring meaning (**regularity**). The degree of regularity in this simulation is only small (the r correlation measure is around 0.2). In general, regularity can be difficult to obtain because to go from a irregular to a regular lexicon, many changes to the lexicon are required, and these changes might involve a decrease in fitness for the individual that adopts them (because non-regular lexicons can also be evolutionary stable states). Moreover, the contribution to the communicative success is small in comparison to the other three properties, because it only plays a role for signals that are radically misperceived. Nevertheless, regularity does emerge in many of the simulations. In the next section I show results

in the deterministic lexicon condition, where simulations run extremely fast. In this condition, many more, and more large-scale experiments could easily be performed, such that I can report quantitative results on the prominence of regularity.

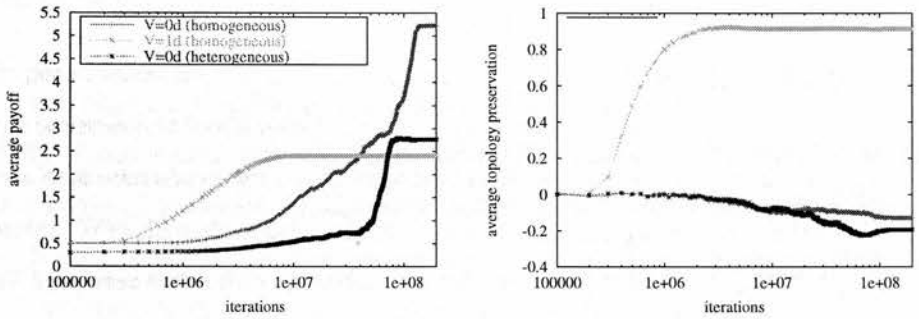
5.7 Local Optimisation of a Deterministic Lexicon

Figure 5.9 shows the average payoff and topology preservation for simulations under 3 different conditions: (i) homogeneous and no topology in the meaning space ("$\mathbf{V} : 0d$"); (ii) homogeneous and $\mathbf{V} : 1d$; (iii) heterogeneous and $\mathbf{V} : 0d$. The results are plotted with a logarithmic x-axis.

A first striking result from these simulations, is that convergence is more than 10 times faster if there is a topology in the meaning space. To understand why, we should first ask why convergence takes so long in the $\mathbf{V} : 0d$ condition. As I discussed above, at any point in the simulation, there is for each meaning a specific signal that has the highest chance of being understood correctly, and similarly for each signal a specific meaning that is its most probable correct interpretation. The optimal behaviour is therefore for all agents the same (assuming a large population). However, the local hill-climbing algorithm that agents use bases its decisions at every step on a sample of just one agent from the population. Stochastic fluctuations will therefore mean that initially (as long as the differences between alternate signals for one meaning, and alternate meanings for one signal are small) different agents will make different changes to their lexicons, and convergence is postponed.

Suppose that, at some point, there is for a particular signal s a specific dominant interpretation m in the population's $\mathbf{R}$ matrices. Now consider the changes the hillclimbing algorithm will favour in the agents' $\mathbf{S}$ matrices. In the $\mathbf{V} : 0d$ condition, a signal is either correctly interpreted or it is not. Unless the random change the hill-climbing considers is signal s, an agent will thus remain at a random other signal for meaning m. If there are 49 different signals, 48 out of 49 iteration of the hill-climbing algorithm do not contribute to convergence. In the $\mathbf{V} : 1d$ condition, in contrast, it does make a difference which of the "wrong" signals is used. Signals that are closer to s, even if they are not equal to it, will give a higher payoff than those that are further a way. Many more of the 49 iterations will now contribute to convergence, if only a little bit.

I suspect a similar mechanism is responsible for another striking result from these simulation. Figure 5.10 shows the average payoff and topology preservation for 60 simulations where the dimensionality of the signal space is varied. In all cases, the payoff reaches high levels (when the signal space is 1d) or intermediate levels (when the signal space is 2d and the overall noise-level is consequently higher because each signal has more neighbours). Importantly, in all cases the



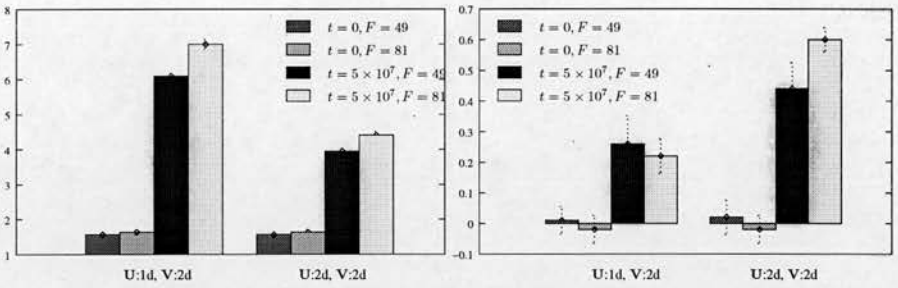
(a) Payoff ($\overline{w}$) (b) Topology Preservation ($\overline{r}$)

Figure 5.9: Average payoff (a) and degree of topology preservation (b) for 2×10^8 iterations under 3 conditions: (1) **V**:0d homogeneous, (2) **V**:1d homogeneous; (3) **V**:0d heterogeneous. The maximum average payoffs that are reached depend on the arbitrary chosen values of the **V** matrices (see also footnote 13); hence, only the shapes of the curves are important. Common parameters are N=400, M=16, F=49, **U**:1d.

topology preservation reaches high levels (when the dimensionalities of meaning and signal space match) or intermediate levels (when the dimensionalities mismatch).

As we have seen above, lexicons do not need to show topology preservation to be evolutionary stable states. Why then, does such a high degree of topology preservation emerge in these simulations? I surmise that a similar mechanism I described above is responsible. Consider again a situation where not all conventions have been established, but a strong association exists between a meaning m and a signal s. What should an agent do to express a neighbouring meaning m'? As long as none of the signals has much chance of being interpreted as m', it pays off for an agent to use a signal s' that is equal or close to s, because interpretation m at least generates some payoff if m' was intended. This intuition – which implies that lexicons with regularity have a larger *basin of attraction* – needs to be worked out more formally, and tested in simulations.

Figure 5.11 shows examples of the **S** and **R** matrices at various stages in the simulations of figure 5.9. Figure 5.12 shows examples of the communication systems in the simulations of figure 5.10. These results (from local optimisation of a deterministic lexicon), show again the same characteristics as before. In addition to *specificity* and *coherence*, *distinctiveness* can be recognised in the **S** matrices, in that the used signals are maximally dissimilar to each other so that they can be easily distinguished (compare figure 5.11a, at the start of the simulation, with 5.11c, at equilibrium). In the **R** matrices, clusters of neighbouring signals all are interpreted as the same meaning. Typically, the most central signal (except at the edges) in such a cluster is the one that is actually used by the **S** matrix (compare figure 5.11c with 5.11d).



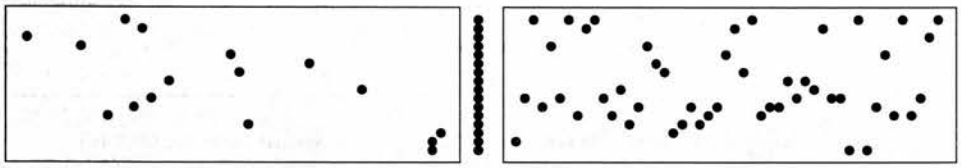
(a) Payoff ($\overline{w}$) (b) Topology preservation ($\overline{r}$)

Figure 5.10: Average payoff (a) and degree of topology preservation (b) at the random initialisation and after 5×10^7 iterations for different parameters. Error-bars indicate standard errors ($\pm\sqrt{\sigma}/N$, where $N = 5$ is the number of simulations with the particular parameters, and σ is the standard deviation). Common parameters are P=400, M=36 and **V**:2d heterogeneous.

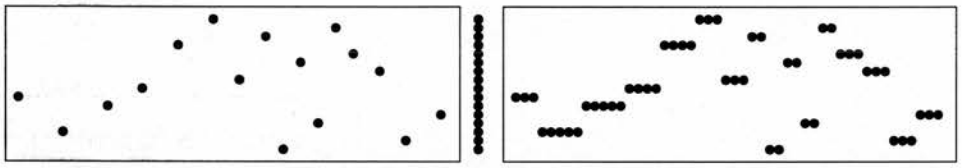
Topology preservation is even more pronounced than in the probabilistic lexicon condition. Again, preservation of topology is not perfect (there is one major irregularity and several minor ones in the signal–meaning mapping of figure 5.11e and f. The topology preservation, according to equation (5.6), is $\overline{r} = 0.915$), but in all simulations performed it is surprisingly high. "Bad" solutions, such as the **S** and **R** of figures 5.11c and d ($\overline{r} = -0.073$), are stable once established in the population, but have a much smaller basin of attraction. In the case of a two-dimensional meaning space, we can draw plots like figures 5.12a-d, which show that the topology is almost perfectly preserved if the dimensionalities of the meaning- and signal-spaces match (5.12a), although it is skewed if different meanings receive very different values (5.12b). But even if the dimensionalities do not match, there is a strong tendency to preserve topology as well as possible (5.12c and d).

When one analyzes the intermediate stages between the random initialization and the equilibrium solutions, it becomes clear that with a heterogeneous **V** valuable meaning-signal pairs get established first, and change little afterwards. This can be seen for instance when comparing figure 5.12d.b with d.c.

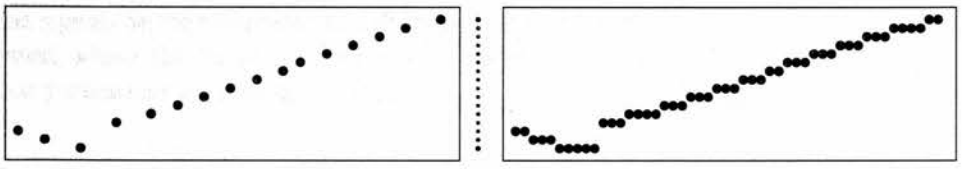
Finally, when the **V** matrix is heterogeneous (figure 5.12b and d), or there is a dimensionality-mismatch (figure 5.12c and d), one can observe that meanings with very low value are sacrificed for the benefit of robust recognition of more valuable meanings (a similar observation was made in Nowak & Krakauer, 1999). These sacrificed meanings "deliberately" get expressed with a signal that will be interpreted with a meaning that is very close. An analogue for this phenomenon in natural language is using a word like "green" to express a color like turquoise, as happens in



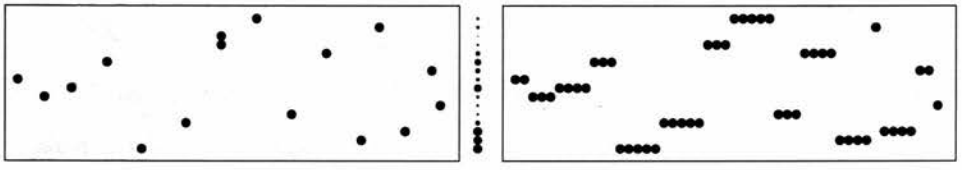
(a) **S**, **U**:1d, **V**:0d homog., t=0 (b) $\mathbf{R}^T$, **U**:1d, **V**:0d homog., t=0



(c) **S**, **U**:1d, **V**:0d homog., t=∞ (d) $\mathbf{R}^T$, **U**:1d, **V**:0d homog., t=∞



(e) **S**, **U**:1d, **V**:1d homog., t=∞ (f) $\mathbf{R}^T$, **U**:1d, **V**:1d homog., t=∞



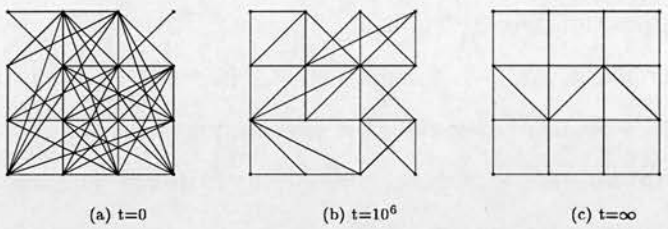
(g) **S**, **U**:1d, **V**:0d heterog., t=∞ (h) $\mathbf{R}^T$, **U**:1d, **V**:0d heterog., t=∞

Figure 5.11: (a)-(h) Examples of $\mathbf{S}$ and $\mathbf{R}$ matrices from the simulations of figure 5.9. For easy comparison, the $\mathbf{R}$ matrices are transposed so that in all matrices meanings differ on the vertical axis, and signals on the horizontal axis. Between the matrices the diagonal values of the $\mathbf{V}$ matrix are plotted, where the diameter of a circle corresponds to value of the corresponding meaning. Common parameters are P=400, M=16, F=49.

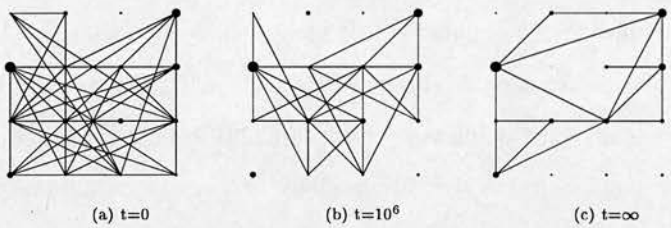
some languages, because a word "turquoise" doesnot exist in the language and a slight misunderstanding is better than no understanding at all.

5.8 Discussion

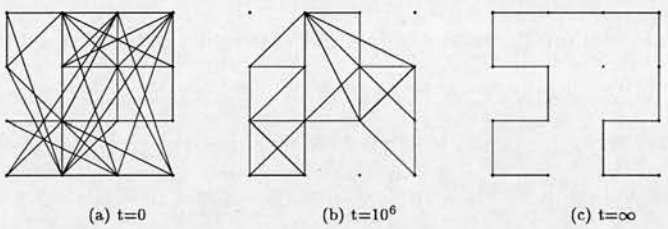
I started this chapter with a brief sketch of compositional semantics in natural language, and some considerations about its evolutionary origins. I agree with researchers like Jackendoff (2002) that compositionality is a fundamental design feature of human languages. I also agree with Jackendoff that in the evolutionary history of language a stage might have existed where language



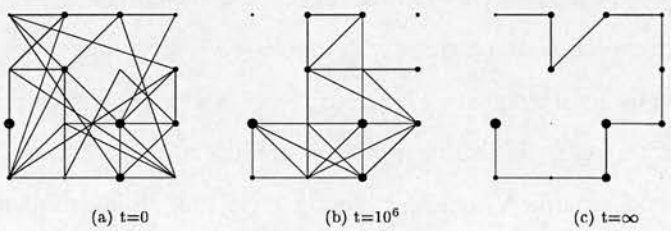
(a) U:2d, V:2d homogeneous



(b) U:2d, V:2d heterogeneous



(c) U:1d, V:2d homogeneous



(d) U:1d, V:2d heterogeneous

Figure 5.12: Topology preservation at equilibrium in 4 simulations with 1d and 2d $\mathbf{U}$ matrices, and homogeneous and heterogeneous 2d $\mathbf{V}$ matrices. Shown are results at initialisation (left column), intermediate time (middle column) and at equilibrium (right column). Nodes are meanings (diameters correspond to value), edges connect neighbours in signal space (several signals can map to a single meaning, such that nodes can have many neighbours; some meanings are not expressed, and the corresponding nodes are not connected). Common parameters are P=400, M=16, F=49.

was productively compositional, but where many of the intricacies of modern syntax, such as hierarchical phrase-structure, were still absent.

Identifying intermediate stages is an important step in constructing a plausible evolutionary scenario; a crucial next step, as I have argued, is to explain the transitions between stages, in this case the transition from a stage without compositionality to a stage with this feature. Moreover, I believe such a scenario must be formalised, such that its internal coherence can be evaluated using techniques from analytic mathematics and computer simulation. Hence, the challenge is to formulate plausible assumptions on the available strategy set and payoff function, and show how compositionality can invade in a population that speaks a language without it.

In section 5.3 I have discussed a number of formal models that take up this challenge, and found that neither is really convincing yet. I have argued that the assumed payoff-function in the model of Nowak & Krakauer (1999) is implausible, because the additional costs of signals of longer duration are not taken into account. If one does consider these constraints, compositional semantics has nothing extra to offer over a phonology that minimises acoustic confusability. I have further argued that the model of Nowak *et al.* (2000) does not deal properly with the invasion of innovations in a population; the model only makes the rather obvious point that a compositional language, once established, allows for generalisation, and hence a greater *average* fitness in the population. Finally, I have discussed Iterated Learning models, like Kirby's (2000), and argued that – for explaining the evolution of compositionality – these models assume too many a priori cognitive abilities; a better model would consider a wider strategy set and consider the selective advantages of compositional versus non-compositional strategies.

In the second part of the chapter I have studied a new model that focuses on a simpler, but related problem: the evolution of a lexicon with **topology preservation**, where similar meanings tend to be expressed by similar signals. I have introduced a formalism to describe the quality of a lexicon, that includes a matrix $\mathbf{U}$ that describes the confusion probabilities of signals depending on their similarity, and a matrix $\mathbf{V}$ that describes the payoff for all intention–interpretation pairs. The *strategy set* the model considers corresponds to all possible choices of $\mathbf{S}$ and $\mathbf{R}$ matrices; the *payoff function* is implicit in the $\mathbf{U}$ and $\mathbf{V}$ matrices. I identified four qualitative properties of the optimal lexicon, and evaluated if they could *invade* in a population with a language without these properties. I found that only lexicons that are specific, coherent and distinctive are evolutionary stable strategies. Moreover, the simulation results suggest that evolutionary stable lexicons that also show topology preservation are much more likely to emerge than those that do not.

Hence, compared to existing models, the model represents progress in meeting the requirements for evolutionary explanations of chapter 2. The model should still be made more formal (with the intuitive arguments for why only lexicons with the listed properties are evolutionary stable states turned into formal proofs), and the "sufficient time" requirement needs to be studied (given that some of the simulations needed 10^8 iterations to converge).

But what exactly is the model explaining? The model shows that with simple assumptions on topologies in the meaning- and signal-spaces, and individual-based optimisation, communication systems can arise that show a structured mapping from meanings to signals. The existence of a topology in the meaning- and signal-space should not be controversial, although it is not obvious how many dimensions these spaces should have, and how payoff and confusability decrease with distance in these spaces. It remains to be shown that a significant degree of topology preservation also emerges if these spaces are of higher dimensionality (and perhaps hierarchical structure).

However, the main limitation of the current model is that topology preservation is not the same thing as compositional semantics. Compositionality does imply that similar meanings are associated with similar sounds: a signal *johnwalks* is similar to both *johnsleeps* and *marywalks*. But topology preservation in natural languages can also be due to sound symbolism, where "Words whose meanings lie close to one another, are likewise accorded similar sounds" (von Humboldt, 1836, p. 74). The cognitive relevance of sound symbolism – such as in examples like *slippery, slimy, sluggish, sloppy, slithery, sleazy* – is controversial, but it is clear that the common sounds in such examples bear no direct semantic content, and the signals are therefore not compositional.

What then, does the model say about the evolution of compositional semantics? The assumption I make is that in a population where a language with some sort of topology preservation is spoken, the fundamental new phenomenon of productive compositionality can more easily evolve. Consider the example of figure 5.12a.c, which is the same system I used in figure 5.3 to explain the use of the two-dimensional meaning- and signal-spaces and the visualisation of the topology preservation. Let us now interpret the horizontal axis of the meaning space as describing agents, ranging from BABY to GRANNY, and the vertical axis as describing actions, ranging from from LIES to RUNS. Meaning 10 in figure 5.13a would then mean something like *"the woman walks"* (the meanings here are thus assumed to be "combinatorial"; recall that I reserved the term "compositionality" in this chapter for a property of the *mapping* between sounds and meanings). Finally, let us interpret the axes of the signal space as describing two components of the signal, for instance the horizontal axis as describing a first sound ranging from /bu/ to /bo/, and the vertical

axis as describing a second sound ranging from /mu/ to /mo/. Signal 18 in figure 5.13b would then correspond with something like /bima/.

	baby	girl	woman	granny
runs	3	7	11	15
walks	2	6	10	14
sits	1	5	9	13
lies	0	4	8	12

(a) The 4×4 meaning space, with the indices of 16 meanings

	/bu/	/bi/	/ba/	/bo/
/mo/	6	13	20	27	34	41	48
	5	12	19	26	33	40	47
/ma/	4	11	18	25	32	39	46
	3	10	17	24	31	38	45
/mi/	2	9	16	23	30	37	44
	1	8	15	22	29	36	43
/mu/	0	7	14	21	28	35	42

(b) The 7×7 signal space, with the indices of 49 signals

42	44	46	48
28	30	32	34
14	16	18	20
0	2	4	6

(c) The 4×4 meaning space, with the indices of 16 corresponding signals, according to the **S** in figure 5.3d.

bomu	bomi	boma	bomo
bamu	bami	bama	bamo
bimu	bimi	bima	bimo
bumu	bumi	buma	bumo

(d) The 4×4 meaning space, with the 16 corresponding signals

Figure 5.13: Topology preservation interpreted as compositional semantics.

The topology preservation that emerged in the simulation of figure 5.12a, under this interpretation of the meaning and signal space, now describes a perfectly compositional system where a "word" like *ba* means WALKS, and a word like *mi* means GIRL. Figure 5.13c (the same as figure 5.3f) and figure 5.13d describe the same meaning space, but the latter represents signals according to the interpretation above. This figure shows how topology preservation can, under some assumptions, be viewed as compositional semantics.

Of course, the model does not model the evolution of *productive* compositionality. That is, it is not concerned with the difficulties with or the benefits of generalising to unseen examples. However, in a population where such a *superficially* compositional language is spoken, a mechanism for productive compositionality does not suffer from the initial selective disadvantages that it would in a population that spoke a language without any structure in the mapping from meanings to signals (due to the problem of coordination). The model therefore represents a possible solution for the difficulties that exist with models like those of Kirby (2000) and Nowak *et al.* (2000).

5.9 Conclusions

The model I have presented in this chapter deals with a lexicon that relates meanings to signals and vice versa. Unlike existing work, I have looked at situations where (i) the payoff of a meaning as an interpretation of a signal, depends on how similar it is to the intended meaning; (ii) where some meanings are more valuable than others; and (iii) where signals can be confused with each other depending on their similarity. I found that the optimal lexicon, as well as the evolutionary stable lexicon, show the following properties: specificity, coherence, distinctiveness, regularity and the sacrificing of meanings with low value.

This model is perhaps interesting in itself, and the measures could potentially be related to empirical observation of communication systems. For the evolution of compositional semantics it offers a possible solution to some of the problems of existing models. Analogous to the model of chapter 4, this model shows a *path of ever increasing fitness* from a non-compositional to a *superficially* compositional language. I suggest that this can be the intermediate step that *productive* compositionality needs to invade in a population.

CHAPTER 6

Hierarchical Phrase-Structure[1]

In this chapter I discuss a third major transition in the evolution of language: the emergence of hierarchical phrase-structure. I first briefly sketch what it is, and then introduce some of the formal models that have been proposed to describe its nature, its acquisition and its evolution. I then present a new model that relaxes some of the simplifying assumptions in existing models.

[1]The work that I describe here builds on joint work with Paulien Hogeweg, which appeared in Zuidema & Hogeweg, 2000 (see appendix C of this thesis), and with Tim O'Donnell, which appeared in O'Donnell & Zuidema, 2004 (see appendix C). Most results described here have appeared in Zuidema, 2003a (see appendix C). All modelling, graphs and text in this chapter are my own.

6.1 Introduction

6.1.1 Phrase-Structure in Natural Language

In the previous chapter I have discussed compositionality in natural language, where the meaning of combinations is a function of the meaning of the parts and *the way they are put together.* The combination of a proper noun (*Mary*) and an intransitive verb (*walks*) is perhaps the simplest example: the meaning of the compound depends on the meaning of the parts. However, the way words and morphemes are combined in most natural language sentences is considerably more complex. Consider the following example sentence:

(6.1) *"The clever mouse enjoys seeing the dog chase the cat".*

Obviously, the meaning of the sentence depends on the meaning of the words, but crucially these meanings need to be combined in a specific order. Thus, *clever* says something about *the mouse*, and not about *the dog* or *the cat.* Similarly, only *the mouse* is enjoying and seeing, and only *the dog* is chasing. The words *clever* and *the mouse* thus need to be first combined with each other, and subsequently with *enjoying* and so-forth. Moreover, if we compare this sentence with for instance *"the mouse enjoys sleeping"*, it is clear that *the clever mouse* and *the mouse*, and *sleeping* and *seeing the dog chase the cat* play the same parts in their respective sentences. Such **phrases** (a word, or a number of words grouped together) can be used in the same positions, and, hence, are of the same **syntactic category**. Other phrases cannot be used in the same positions; they are of a different syntactic category.

The fundamental observation is that underlying a sentence like example 6.1 is a level of organisation that we can call **phrase-structure** (Chomsky, 1955, 1957; Higginbotham, 1997). Phrases can be further combined into larger phrases, i.e. the structure is *hierarchical*, and such larger phrases might be of the same syntactic category as one of the smaller phrases they contain, i.e. the structure is *recursive.* All combinations are guided by rules of combination that regulate which phrases of which category can be combined into larger phrases. The phrase-structure of a sentence can be represented with brackets, or more graphically as a tree. We can assign the following structure to the example sentence:

[[The [clever mouse]] [enjoys [seeing [[the dog] [chase [the cat]]]]]]. (6.2)

If we add linguistic category labels to each of the phrases, this becomes a cumbersome formula, which is more clearly represented as the tree in figure 6.1 (here difficulties with inflection, agreement and articles are ignored).

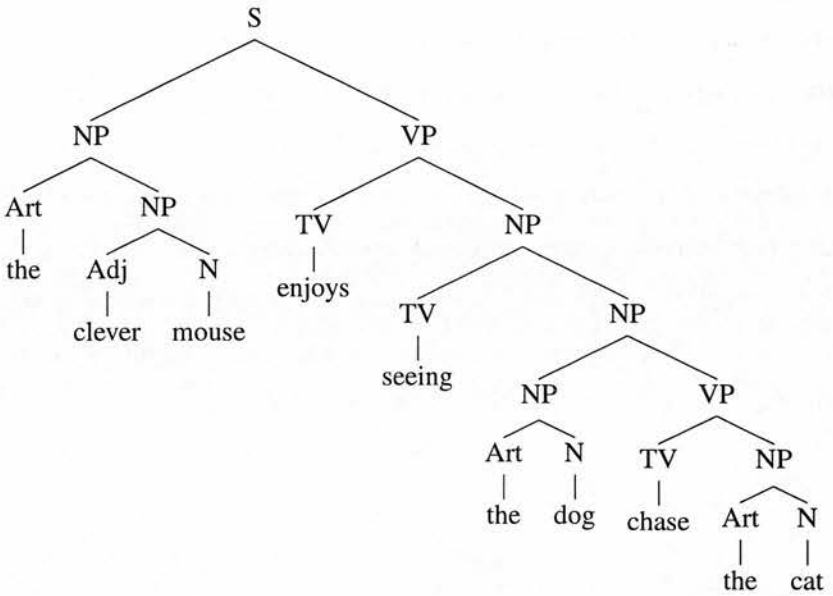


Figure 6.1: The conventional tree representation of the hierarchical phrase-structure of the example sentence.

Since Chomsky (1957), recursive, hierarchical phrase-structure (henceforth, "phrase-structure") has been widely recognised as a crucial design feature of human language. Jackendoff (1999, 2002) lists phrase-structure (without the subtleties of syntax in modern languages, such as function words, agreement and case marking) as one of the major innovations in the evolution of language. Current linguistic theories differ in whether they consider phrase-structure a primitive or an emergent property of resolving the semantic and syntactic dependencies in a sentence (Rambow & Joshi, 1994). Nevertheless, there is consensus that formalisms for describing natural language syntax need to account for the hierarchical and recursive structure of sentences. Context-free phrase-structure grammars are the archetype formalism that can deal with these features, but many other adequate formalisms exist. However, formalisms such as Markov processes (probabilistic finite-state grammars) or schema's with fixed slots (1st, 2nd, 3d word etc. in the sentence), which occasionally emerge in debates about domain-specificity and innateness of language, fail this requirement. In natural languages, phrases can, at any point in a sentence, be blown-up to arbitrary length (e.g. replace *"the clever mouse"* with *"the clever mouse, who all children in the world love and admire,"*). The inadequate formalisms would need a new path or schema for each extension, without a *systematic relation* between subsequent versions of the sentence[2].

In contrast, **context-free phrase-structure grammars** can deal with such *long-distance dependencies*, and recursive, hierarchical phrase-structure in general. Context-free grammars are

[2]Researchers proposing such formalisms often argue that "language really is finite" (e.g. Reich, 1969); but the issue is not with (in)finiteness, but with a systematic relation between grammatical sentences.

rewriting grammars, and thus specified by two sets of symbols, the terminal symbols V_{te} and the non-terminal symbols V_{nt} and a set of production rules (see table 3.1 in chapter 3). The set of strings that a rewriting grammar can generate is called a *language*; context-free grammars can generate languages from the *class* of context-free languages. Informally, the context-freeness implies that the production rules are restricted to those of the form $A \mapsto \sigma$, where A is a single, non-terminal symbol ($A \in V_{nt}$), and σ is a string of any number of terminal or nonterminal symbols ($\sigma \in (V_{nt} \cup V_{te})^*$). A sentence like example 6.1 is *derived* by starting with a start-symbol S, and replacing it with the symbols NP and VP by applying a rule $S \mapsto NP\ VP$. The tree in figure 6.1 shows all the subsequent applications of rules, all of which are context-free, necessary to finally produce the whole sentence.

With the analysis of the power of different formalisms, Chomsky established a hierarchy of formal languages that is now termed the **Chomsky Hierarchy** (introduced in chapter 3). Finite-state languages in that hierarchy are languages that can be recognised by rewriting grammars with more restrictions than context-free (there is only a single non-terminal on the right-hand side of rules, and all terminal symbols occur on the right side of that non-terminal); context-sensitive languages can be recognised by grammars with fewer restrictions (rules may be of the form $vAw \mapsto v\sigma w$, where A and σ are defined as before, but $v, w \in (V_{nt} \cup V_{te})^*$ represent a context that is a necessary condition for application of the rule).

A fundamental task for theoretical linguistics has been to locate adequate formalisms for natural language syntax on that hierarchy. Chomsky (1957) argued that even context-free grammars are not powerful enough to model some frequent syntactic phenomena, and proposed "transformations" as a solution. In the early eighties it was shown that Chomsky's original arguments did not hold (Gazdar, 1981), and, in most linguistic frameworks, the traditional transformations were abandoned again. Only a few years later, it emerged that there are in fact syntactic phenomena in natural languages – although different from Chomsky's examples – that place them outside the class of context-free languages (Huybrechts, 1984; Shieber, 1985). It is now believed that for adequately modelling natural languages, the *weak generative capacity* of a formalism needs to be slightly more than context-free.

However, it is also clear that not all context-sensitive languages can be natural languages, because that class includes languages that are completely dysfunctional for communication (for instance, languages with only prohibitively long sentences). The class of possible natural languages is thus a different set, that is likely to be a subset of the context-sensitive languages (and

probably disjoint from the context-free, and hence, finite-state languages[3]) but that is constrained in many ways that have nothing to do with the Chomsky Hierarchy. This is sketched in figure 6.2. The *class* of possible human languages is sometimes termed "Universal Grammar" (e.g. Nowak *et al.*, 2002), although that term is more commonly used to describe the universal, innate component of natural languages.

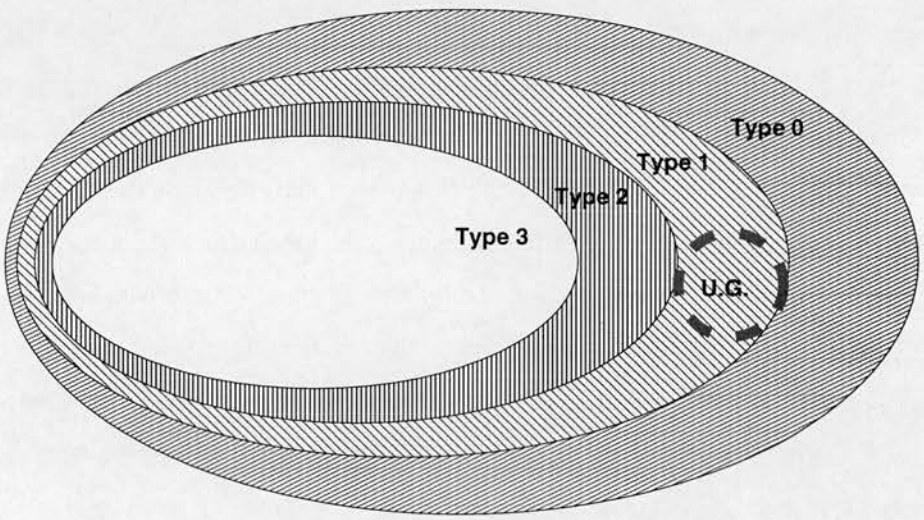


Figure 6.2: The four main classes of languages from the Chomsky Hierarchy and the class of possible natural languages (UG). Type 3 languages are those recognised by finite-state grammars; type 2 those recognised by context-free grammars; type 1 those recognised by context-sensitive grammars; type 0 those recognised by any rewriting grammar (that is, by any computable function / Turing-machine). Note that $\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$. That is, the class of type 3 languages is a proper subset of the class of type 2 languages and so forth.

Using empirical observations on linguistic diversity and language acquisition and use, linguists have tried to identify the relevant constraints and to find formalisms that account for such constraints in the most natural way possible. Joshi (1985) proposed, based on his work on "Tree-Adjoining Grammars", that the subclass of context-sensitive languages that are good models of natural language, can be characterised as the class of *mildly context-sensitive* languages. Such languages have a number of special properties, including being parsable in polynomial time and the so-called "constant growth" property. Since several other popular formalisms have been shown or conjectured to be mildly context-sensitive (Joshi *et al.*, 1991), some consensus has emerged about the upper and lower bound on the power of grammar formalisms.

[3]Note, however, that it has not been established that *all* natural languages fall outside the context-free or even the finite-state languages, nor that no (currently unknown) syntactic constructions exist in *some* language that go beyond (mildly) context-sensitive power. Hence, the class of possible natural languages could intersect with all main classes of the Chomsky Hierarchy. Most current theories of syntax, however, assume that the computational procedures underlying all human languages are very similar; if the semantics in one human language requires trans-context free power, then it is likely, according to these theories, that all human languages do.

Unfortunately, otherwise much disagreement remains in the field about the appropriate formalisms and the nature of these constraints, and a great many alternative frameworks for describing syntax exist, each with many practitioners. I will not attempt to review these frameworks, but they include Government & Binding / Principles & Parameters (GB/PP, Chomsky, 1981), Head-driven Phrase-Structure Grammar (HPSG, Pollard & Sag, 1994), Combinatory Categorial Grammar (CCG, Steedman, 2000; Steedman & Baldridge, 2003), Optimality Theory (OT, Prince & Smolensky, 2004), Lexical-Functional Grammar (LFG, Kaplan & Bresnan, 1982), Tree Adjoining Grammars (TAG, Joshi *et al.*, 1991), the Minimalist Program (MP, Chomsky, 1995) and Construction Grammar (Kay & Fillmore, 1999; Goldberg, 1995). Common themes can be identified in recent developments in these different frameworks. Other than mild context-sensitivity, these include lexicalisation (where productive rules are always associated with specific words), heterogeneity, redundancy and stochasticity. Nevertheless, the differences between the frameworks – both in methodology and in content – are enormous, and a major obstacle for pursuing interdisciplinary work on the psychology, biology or indeed the evolution of language.

6.1.2 Evolution of hierarchical phrase-structure

Despite the many controversies in linguistics, there seems to be a consensus that natural languages exhibit recursive, hierarchical phrase-structure. It is clear that this feature poses requirements on the cognitive abilities of language users. They need to be able to produce and interpret sentences with that structure. Moreover, infants need to acquire the syntax of their native language from observations of the use of language around them, without much or any explicit instruction. With language being such a salient behaviour of humans, the origins of these abilities in humans, and these patterns in natural languages, are a fundamental question for both evolutionary biology and cognitive science.

However, the *poly-paradigmatic* state of linguistics presents evolutionists with a difficult problem: how can we say anything sensible about the origins of phrase-structure and Universal Grammar, if we cannot even agree on what it is and how we should describe it? In particular, even if we agree on a description of hierarchical phrase-structure, how do we decide on a reasonable strategy-set and payoff function that we need for an evolutionary scenario?

Many linguists have simply resisted speculating about these issues. Noam Chomsky, notably, has dismissed such speculations as untestable stories (e.g. Chomsky, 2002). Many non-linguists, on the other hand, have simply ignored the complexities of syntax, and have focused on speech and compositionality instead, apparently assuming that syntax would simply follow (as the rare linguists concerned with evolution complain, e.g. Newmeyer, 2003; Bickerton, 2003b; Hurford,

2002b). Yet others, including Pinker & Bloom (1990) and Jackendoff (2002) have taken up the challenge and provided some intuition for selective advantages of grammatical constructs. However, these verbal accounts have remained so much underspecified that it is difficult to even start constructing the assumed strategy set, payoff function and initial selective advantage.

In this chapter I will evaluate a number of more or less formal approaches to this issue. I will first discuss the models of Batali (2002) and Kirby (2002a) and a number of related models. These authors leave hardly any role for natural selection (in a sense, they consider a strategy-set in the evolution of grammatical language that includes just one learning strategy). Batali and Kirby view hierarchical phrase-structure as an emergent property of the negotiation or iterated learning of a communication system in a population of agents. I will argue, as I did in previous chapters, that we need to evaluate the fitness consequences of different outcomes of such self-organising processes. However, I will also argue that these models show that the evolution of recursive, hierarchical phrase-structure differs in important ways from the evolutionary problems considered in chapters 4 and 5.

Secondly, I will discuss a number of studies (Hashimoto & Ikegami, 1996; Nowak *et al.*, 2002; Fitch & Hauser, 2004) that implicitly or explicitly use classes from the Chomsky Hierarchy as the strategy set. This is an attractive approach, because the Chomsky Hierarchy offers some well-understood concepts and tools, and the location of natural language on the Hierarchy is one of the few topics in linguistics for which a broad consensus exists. Nevertheless, I will argue that the Hierarchy is not suitable to serve as a strategy set, because it is not fine-grained enough and because the different class boundaries in the hierarchy have no natural biological interpretation.

Thirdly, I will discuss a model by Nowak *et al.* (2001). This model is based on what we can call the *Uniformity Assumption*: the idea that all possible languages are of equal quality and equally likely to be the target of learning. This model is interesting because it is elegant and ambitious. However, I will argue that the crucial dependence on the Uniformity Assumption ultimately makes the model of Nowak et al. of limited use.

In the rest of this chapter I will present a new model that further illustrates the difficulties with the Uniformity Assumption in language evolution models, and highlights the interactions between cultural and biological evolution. I will argue, based on this model, that a better understanding of this interaction is crucial for understanding the origins of phrase-structure.

6.2 Related Work

6.2.1 Cultural Evolution in Expression–Induction models

Batali (2002) presents an explanation for the origins of phrase-structure in natural languages

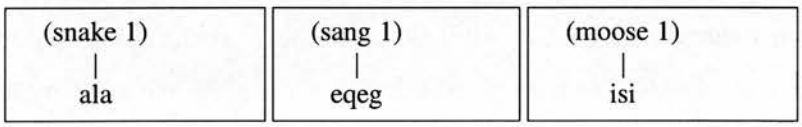
that does not involve biological evolution. Batali is interested in the properties of a language "negotiated" between agents in a population. The agents in his simulation come equipped with the ability to represent grammatical structures, to produce and interpret sentences, and to learn from experience (induction)[4]. In every step in the simulation, a random speaker is selected from the population and confronted with a logical formula from a predefined meaning space. The speaker produces an utterance, and the hearer receives both the utterance and the meaning and updates her memory. The model is an instance of what Hurford (2002a) calls "Expression–Induction models", where structure emerges in a cycle of expression of I-language as E-language, and induction of I-language from E-language.

In the model, knowledge of words and grammar is encoded in a collection of "exemplars", each with an associated cost. When an agent in the role of speaker is presented with a meaning, she searches for the cheapest way to express it. In the initial phase of the simulation, there is no common language and she simply generates a random string of characters, even though a high cost (proportional to the number of symbols in the signal and the number of predicates in the meaning) is associated with this operation. When agents, in the role of hearers, receive an unrecognised form–meaning pair, they simply store it as a holistic exemplar (a tree of depth 1), with associated initial cost $c_0 = 1.0$.

After a number of cycles, agents will have stored a number of different exemplars. With more and more exemplars stored, it becomes increasingly likely that the cheapest way to express a given meaning is by reusing an existing exemplar, or by combining or modifying exemplars. Two exemplars A and B, when combined, yield an exemplar E that is a tree, with as root the combination of the meanings of A and B (with the arguments possibly renamed), and as daughters the two exemplars (see figure 6.3). An exemplar can be modified by replacing a subtree. The costs of a combination of exemplars is the sum of the costs of the parts, plus an additional cost for modifying and combining them. This way, tree-structured exemplars are created; nodes in these trees are labelled with semantic information (predicate logic expressions, rather than conventional part of speech tags).

The "best" phrase-structure of a given sentence in Batali's model, is the one that has the lowest total costs. Costs in Batali's model thus play the same role as probabilities in the probabilistic grammar formalisms used in computational linguistics (e.g. Manning & Schütze, 1999). In its reliance on storage of many exemplars, the model is reminiscent of the Data-Oriented Parsing model from that field (Bod, 1998).

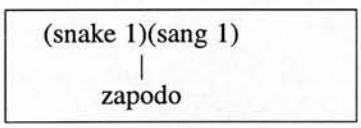
[4]I use the terms "learning" and "induction" interchangeably and in a very broad sense, that includes any change in knowledge and behaviour in response to environmental input.



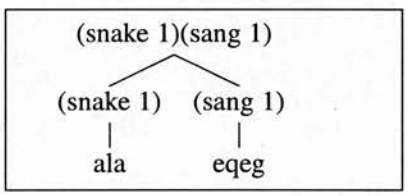
(a) Exemplar A with cost k_A, pairing the signal *"ala"* with the meaning *"object 1 is a snake"*.

(b) Exemplar B with cost k_B, pairing the signal *"eqeg"* with the meaning *"object 1 sang"*.

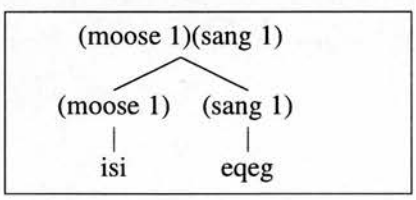
(c) Exemplar C with cost k_C, pairing the signal *"isi"* with the meaning *"object 1 is a moose"*.



(d) Exemplar D, resulting from the "invention" operation, with cost $k_D = k_{invention}$, pairing the signal *"zapodo"* with the meaning *"the snake sang"*.



(e) Exemplar $E = A \circ B$, resulting from combining existing exemplars, with cost $k_E = k_A + k_B + k_{combine}$, pairing the signal *"alaeqeg"* with the meaning *"the snake sang"*.



(f) Exemplar $F = E \leftarrow C$ with cost $k_F = k_E + k_C + k_{modify}$, pairing the signal *"isieqeg"* with the meaning *"the moose sang"*.

Figure 6.3: Exemplars and combinations of exemplars in Batali's (2002) model.

When an exemplar is reused and leads to a successful communication (that is, the speaker's intention and the hearer's interpretation are identical), its cost goes down. Therefore, exemplars that prove useful in different combinations will be favoured, giving rise to the *cultural evolution* of language structure[5]. Batali finds that the competition between exemplars leads to recursive and compositional grammars, with which the agents can communicate about many more meanings than they have exemplars. Batali estimates that the languages negotiated in the final stages of the simulations can accurately convey 2.3×10^{13} different meanings (with a communicative accuracy of around 98%), after fewer than 10 thousand learning observations and with several hundreds of exemplars stored (unused exemplars are removed after a few cycles). Crucially, the languages in these simulations, as in the Iterated Learning Model (Kirby, 2000), change and adapt to the bias of the learning algorithm. The emerging languages share some important characteristics with natural language including compositionality, phrase-structure and recursion, as well as specific features such as agreement, reflexives and function words.

Steels (2004) presents a similar model, with a population of agents that negotiate grammatical rules to express predicate-logic formulae. Like Batali, Steels' rules can associate a single word or chunk of words with a single meaning, or associate with complex meanings. Also like Batali, Steels associates a score with each of the rules in the emerging grammars. These scores go up and down with successful or unsuccessful use, and regulate the choice from a set of alternative parses.

The model differs from Batali's in a numbers of ways. First, Steels' formalism is much more complex and, among other things, allows for both flexible and fixed constituent order rules. Such rules are expressed as optional constraints. E.g., *precedes*(x, y) expresses the constraint that the yield of constituent x must precede the yield of constituent y. The formalism decouples *immediate dominance* from *linear precedence* (Gazdar & Pullum, 1981). Second, in Steels' "constructivist" approach, a speaker uses herself and her knowledge of language as a model of the hearer. When given a meaning that cannot be readily expressed with existing rules, new rules are created to generate an utterance that the speaker herself would interpret correctly. Finally, in Steels' model the meanings are grounded in an actual machine vision system that generates predicate-logic expressions from video input, which leads to occasional confusion about the topic of a conversation. It is not clear, however, how these expressions differ qualitatively from the predefined meaning space of Batali.

[5]I will use the term "cultural evolution" here in a broad sense, without specifying the reproducers, replicators and selection pressures as discussed in chapter 2. It would be interesting to work out in detail the analogy between biological evolution and the dynamics in the models of Batali and others, but for now it suffices to note that in these models structure emerges over time (hence, "evolution") in a process where learners learn from learners (hence, "cultural").

Steels' ideas about the learning mechanism are interesting, but unfortunately difficult to evaluate because the algorithm is not precisely defined, the relation with existing models ignored[6] and very few results have been reported. Both Steels and Batali have chosen to work with idiosyncratic formalisms instead of some of the well understood formalisms from theoretical linguistics, presumably out of disagreement with the "nativist" theories of language, in the context of which most of these formalisms were originally developed. Nevertheless, I strongly suspect their formalisms are in fact formally equivalent to some member of the family of stochastic tree grammars (Joshi & Sakar, 2003); for comparison with other work, it would be better to design learning models that work with well-understood formalisms instead.

Hurford (2000) and Kirby (2002a) present related models that do use such well-understood formalisms. Both are versions of the iterated learning model discussed in chapter 5 (Kirby, 2000). Recall that in that earlier model, individuals can produce and interpret sentences, and have a language acquisition procedure to learn a context-free grammar from each other. The model considers the transmission of language from generation to generation, where each generation is represented with just a single agent (or a small number of agents, as in Hurford's model). At every step the *parent* presents a relatively small number of examples of form–meaning pairs to the *child* (the very first parent creates random strings for each of the meanings it wants to express). The child then uses these examples to induce her own grammar. In the next iteration the child becomes the parent, and a new individual becomes the child. The process is repeated many times. In the iterated transmission steps the language becomes easier and easier to learn, because the language adapts to the learning algorithm by becoming more and more structured. Note that knowledge of language is transmitted *vertically* from generation to generation, unlike the models of Batali and Steels where the language is negotiated *horizontally* with other members of a fixed population.

In Kirby (2000), there was a finite number of possible meanings – all combinations of agents, actions and patients. Both Hurford (2000) and Kirby (2002a) use a predicate logic based semantics that has recursive structure. They both find – using different learning algorithms – that the emerging grammars show recursive, hierarchical phrase-structure. Not only simple compositionality, but also the phrase-structure of natural languages could be the result of a cultural selection pressure for increased learnability. Interestingly, the necessary constraints for learnability need not evolve as a separate restrictive Universal Grammar, as in some "nativist" theories of language acquisition, but follow logically from the fact that a child only needs to learn languages that have

[6]Steels (2004), when introducing the syntactic formalism, makes no reference to existing linguistic formalisms (except construction grammar) despite obvious parallels, nor does he, when introducing the learning procedure refer to any existing work on grammar induction, including even Batali's.

been learnt by previous generations. This point will be worked out later in this chapter and in the next.

The use of more or less standard formalisms makes these models easier to understand, and reveals a number of strong assumptions, including a recursive, hierarchically structured meaning space and "innate" procedures for searching and combining the units of language. Similar assumptions were made in the models of Batali and Steels, but because of the unconventional formalisms and the lack of details, it remains difficult to describe and evaluate these assumptions. In Kirby's model each generation consists of just a single agent and the model considers a large meaning space. Hurford's model considers a small population of four agents per generation, and only a small meaning space. An important question is how much the results depend on the specific choices for modelling learning, meaning, grammar, the interaction between agents and for the population size. I am not aware of any subsequent work on Hurford's model, but Smith & Hurford (2003) reimplemented Kirby's model and studied how well it fares in a larger population, where agents learn from multiple "cultural" parents and from peers. They report that similar results could be obtained, but only with a very careful choice of parameters.

More work remains to clarify the relation between the different models and to identify the necessary and sufficient conditions for recursive, hierarchical phrase-structure to emerge. If the results are confirmed in subsequent work this would constitute a compelling explanation for the (proximate) origins of phrase-structure in human languages: in a population of agents with cognitive abilities and communicative intentions as in this model, it will emerge as the result of the negotiation of a language in the population. The models, however, do not explain the (ultimate) origins of these cognitive abilities. Why do agents have the production, interpretation and acquisition procedures that they have? It is possible, of course, that these abilities are accidental properties of the human brain, that evolved under selection pressures independent from language. It is difficult to assess the plausibility of that assumption. In the models, a process of cultural evolution produces languages with interesting properties; as I argued in previous chapters, the next step is to study the fitness consequences of such processes. Results such as those of Batali, Steels, Hurford and Kirby would be much strengthened if one could show that the learning algorithm used is one from a family of "natural" algorithms, and that within that family there is a path of ever increasing fitness towards it.

I have reviewed these Expression–Induction models here at some length because I believe they bring an important lesson for modelling the evolution of syntax. These models show – as does the model developed later in this chapter – that, for each learning algorithm A, it is useful

to distinguish between the set of languages $\mathcal{R}_A$ that can be *represented* by the formalism used by A, the set $\mathcal{L}_A$ that can be *learnt* by A, and the set $\mathcal{I}_A$ that are stable outcomes of a process of cultural evolution that results from repeated application of A. In this chapter, I will explore the implications of this distinction for scenarios of the evolution of phrase-structure.

The difficulties with formulating relevant models of the evolution of hierarchical phrase-structure arise, in part, from the fact that human languages are both the result and the object of a learning process. When we talk about the biological evolution of language, we really talk about the evolution of the learning mechanisms. In the case of phonology and compositionality, it seemed reasonable to describe the strategy set for evolution in terms of the end results of a learning process. Implicit in that decision was the assumption that the signal space considered in chapter 4 and the signal–meaning mappings in chapter 5 were reasonably close to the actual strategy sets of learning mechanisms available for evolution.

In the case of phrase-structure it is much harder to choose a reasonable strategy set, because the learning problem is much more difficult. Given a fully defined learning algorithm like Batali's (2002) or Kirby's (2002), it is extremely difficult to describe (i) the set of languages that it can learn ($\mathcal{L}_A$), (ii) the languages that would result from cultural evolution in a population ($\mathcal{I}_A$), (iii) the changes to these sets when we make a small change in the learning algorithm. Yet, we need to characterise these sets to assess the fitness consequences of changes to the algorithm. I will now discuss accounts of the evolution of phrase-structure based on notions from the Chomsky Hierarchy, and argue that the failure to account for learning and cultural evolution is where they go wrong.

6.2.2 *Natural Selection & the Chomsky Hierarchy*

When one, unlike Batali and others, tries to give an account of the origins of phrase-structure that does involve natural selection, one immediately faces the problem of formulating a plausible strategy set and payoff function. Given the consensus on characterising natural languages in terms of the Chomsky Hierarchy, and the mathematical sophistication of such characterisations, it seems an attractive proposal to define the strategies and payoffs in the evolution of syntactic language in the same terms. For instance, Nowak *et al.* (2002), and similarly Komarova & Nowak (2003) review formal approaches to the evolution of language. They present the Chomsky Hierarchy for describing language, along with (statistical) learning theory for describing language acquisition and the replicator equations for describing evolution. They conclude that "these approaches need to be combined", but do not attempt such an integration.

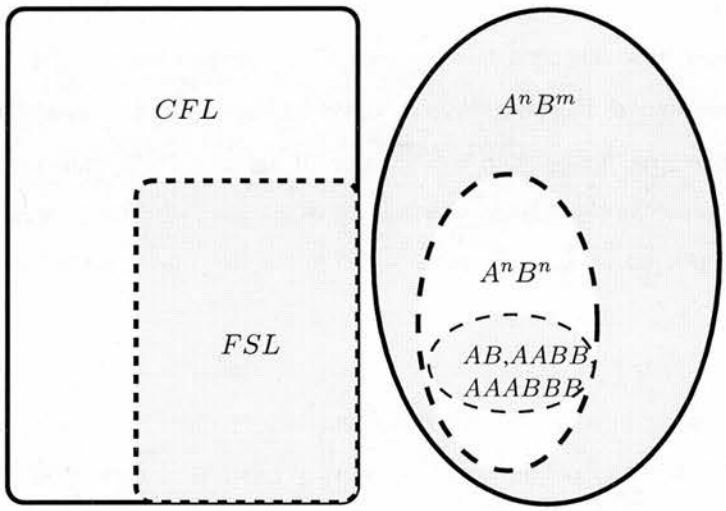
I believe a serious attempt to integrate these formal frameworks would soon show that the Chomsky Hierarchy is in fact of little help in understanding the origins of syntax. As I briefly discussed in chapter 3, the classes of the Chomsky Hierarchy are too coarse. When considering the architectural constraints on the neural hardware, it seems that complexity in terms of the hierarchy is extremely easy to get (Siegelmann & Sontag, 1991; Wolfram, 2002). When considering the difficulties of parsing or learning, it seems the relevant distinctions cut through all main classes of hierarchy (Barton & Berwick, 1987; Gold, 1967). Nevertheless, the idea of evolution having "climbed the Chomsky Hierarchy" is implicit in many accounts. Two papers that have worked this out in some more detail are Fitch & Hauser (2004) and Hashimoto & Ikegami (1996).

Fitch & Hauser (2004) attempt to show that in evolution, human processing capabilities have crossed the boundary between finite-state and context-free languages, whereas those of tamarins (and by assumption other non-human primates) have not. They presented human and monkey subjects in the experiment with strings drawn from the finite-state language $(ab)^n$ (with for instance the strings $abab$ and $ababab$) or from the context-free language $a^n b^n$ (with the strings $aabb$ and $aaabbb$). The a's and b's are monosyllabic sounds produced by a human male and female respectively. The n in the experiments is limited to values up to $n \leq 3$.

The subjects were trained on samples from one language, and then tested on whether they can distinguish samples from this language from the other. Fitch & Hauser found that humans easily pass this test: when trained on either the finite-state or the context-free language, they reject samples from the other. Tamarins, on the other hand, fail the test in the context-free condition. When trained on the context-free language, they accept both the context-free and finite-state test samples.

These results are intriguing. At first sight they support the idea that humans have, in their evolution, moved up to a different level of the Chomsky Hierarchy, and some researchers have already enthusiastically hailed this conclusion (Friederici, 2004). However, there is a number of difficulties with this analysis. First, in experimental formal language theory (O'Donnell, 2004) an inherent methodological difficulty is that although formal languages are typically infinite, experiments necessarily work with finite data. To support the conclusion above, one needs to show that the language human subjects acquire really is a context-free language, and not the finite subset that they see during training (see figure 6.4). This can be assessed by testing whether subjects *generalise* to samples with a larger n than they were trained on. In the supplementary material to Fitch & Hauser (2004), the authors report that this experiment was performed and that the human

subjects indeed seem to have acquired the context-free language, but a recent replication disputes this claim (Perruchet & Rey, 2004).



(a) The *class* of context-free languages (CFL) is a superset of the *class* of finite-state languages (FSL).

(b) The context-free *language* A^nB^n is a subset of the FSL A^nB^m, and a superset of the FSL $\{AB, AABB, AAABBB\}$.

Figure 6.4: In formal language theory, a language is a possibly infinite set of strings over some alphabet, such as the languages A^nB^m, A^nB^n or $\{AB, AABB, AAABBB\}$. A class of languages is a possibly infinite set of languages. Two important classes are the finite-state languages (FSL), such as the language A^nB^m, and the context-free languages (CFL), such as the language A^nB^n. Crucially, the class of finite-state languages is contained in the class of context-free languages (fig. a). The problem for experimental formal language theory is that every context-free language is a *subset* of some finite-state language, as well as a *superset* of many other finite-state languages (fig. b). To make the argument that humans, but not monkeys, can process a language that is outside the class of finite-state languages, one needs to show that every possible way of distinguishing the grammatical from the ungrammatical utterances must involve hypothesising some language outside the set of the finite-state languages. In the case of the Fitch and Hauser experiments this involves, at the very least, ruling out the possibility that what has been learnt is either a finite subset of a non-finite state language (by testing for generalisation to strings with $n > 3$) or a finite-state superset of such a language (by testing for the ungrammaticality of strings with $m \neq n$).

The second, and more interesting problem is that there exists a finite-state grammar a^nb^m that accepts all context-free samples in the experiment, and rejects all finite-state samples (if $n > 1$). That is, the finite-state language a^nb^m properly contains a^nb^n but not $(ab)^n$. Even if one could show that the human subjects did not acquire the finite-state superlanguage[7], the question remains: if tamarins can learn finite-state languages, why haven't they acquired a^nb^m?

[7]Tecumseh Fitch, p.c., reports that this has also been established experimentally.

The fact that they haven't suggests the relevant constraints on learning in this experiment, and the differences between humans and monkeys, are not captured by notions from the Chomsky Hierarchy[8].

A very different study that considers the applicability of the Chomsky Hierarchy to language evolution is Hashimoto & Ikegami (1996). These authors present a simulation model of the evolution of phrase-structure rules in a population of agents. The agents have a fixed, innate context-free grammar, with which they (in the role of speaker) generate strings, and (in the role of hearer) parse strings received from speakers. The fitness of an agent in the model depends, through a rather complicated pay-off function, on the number of successful interactions she has been involved in, either as a speaker or as a hearer, and the length and novelty of the strings produced or received. After a number of interactions in which fitness is assessed, agents produce offspring proportional to their fitness and die. Every agent of the next generation inherits the grammar of a single parent. With fixed probabilities, some mutations occur that add or delete a random rule, or add, delete or substitute a random symbol in a rule.

Of course, context-free grammars can model recursive, hierarchical phrase-structure, but they can also model a simple (finite) "lexical" strategy (see figure 3.4) or (finite-state) tail-recursion. For instance, a grammar $\{\mathtt{S} \mapsto \textit{the cat fears the dog}, \mathtt{S} \mapsto \textit{the dog fears the cat}\}$ generates two sentences holistically, without assigning any phrase-structure to them. At the start of the simulations agents are initialised with just one lexical rule in their grammar. Lexical grammars need one rule for each sentence they generate; combinatorial and recursive grammars, in contrast, can generate many more sentences than they have rules. Because there is no limit on the number of rules, both strategies could in principal generate all possible strings in the finite domain that was used. However, in the mutation scheme used, at most one rule is added at a time. Expressiveness (measured as the number of distinct strings of some finite maximum length) grows much faster with grammar size using a syntactic strategy, and, under the parameter settings considered, syntactic agents can therefore out-compete non-syntactic ones.

Hashimoto & Ikegami make an unconvincing attempt to discuss these results as the initial phase of a climb in the Chomsky Hierarchy. Indeed, the finite grammars used at initialisation are finite-state, whereas the grammars in final stages are not. However, that fact appears accidental, rather than saying anything substantial about the evolutionary dynamics in the simulation. First, with the mutation scheme used, even random drift for a small number of generations would

[8]The fact that there exists a finite-state language that the tamarins cannot learn is not surprising of course. The point here is that the monkeys might not acquire context-free $a^n b^n$ for the very same reasons as why they do not acquire finite-state $a^n b^m$.

take the grammars out of the finite-state class. For instance, finite-state $\{S \mapsto ab\}$ is only two mutations away from context-free $\{S \mapsto ab, S \mapsto aSb\}$. Second, whether or not a grammar is finite-state or not, appears to have very little to do with fitness. The fitness scheme used only considers the number of different strings a grammar can produce and parse. Now observe that a finite-state grammar $\{S \mapsto a|b|aS|bS\}$ is maximally expressive (for $V_{te} = \{a, b\}$), and much more so than for instance context-free grammar $\{S \mapsto ab|aSb\}$.

Note that in this model a grammar that accepts every possible string has maximum fitness, whereas in natural language the whole notion of "grammaticality" implies that many possible strings are in fact ungrammatical. Despite this unrealistic feature, Hashimoto & Ikegami's model fitness scheme leads to some interesting observations. The payoff function has some counterintuitive and interesting consequences due to the fact that it is not a fixed measure, but depends on the kind of grammars that are present in the population. For instance, they find that the most expressive agents are not necessarily the most successful, because they are poorly understood by others in the population, and that a score for *not being recognised* accelerates the evolution of complex phrase-structure.

In my own previous work (Zuidema, 2000; Zuidema & Hogeweg, 2000), I have reimplemented, simplified and extended the Hashimoto–Ikegami model. This new model still uses context-free grammars as the strategy set, and the number of strings in an agent's language that are shared with others as the basis for the payoff function. I found that the reason "syntax" evolved in the original model, is that the authors used a rather arbitrary scoring scheme, where there is a significant payoff for agents that produce novel and complex strings, *independent of whether they are understood or not* (i.e. a selection pressure that is, in part, frequency-independent).

When this unrealistic scoring scheme is replaced by a number of simple payoff functions, an interesting paradox emerges (Zuidema & Hogeweg, 2000). If both speakers and hearers benefit from successful communication (the **mutual benefit condition**), every linguistic innovation represents an initial selective disadvantage, because it leads to increased confusion. In this condition, complex syntax does not evolve[9]. On the other hand, if only hearers benefit (the **hearer benefit condition**), the willingness to speak is lost, because only the speaker's competitors benefit from her speaking. In the first version of the model, where production is enforced, this shows itself in the development of excessively complex grammars, which make the produced strings often impossible to parse with the parsing procedure used. In a later version of the model, a parameter

[9] Although it would be stable once it got established, because it is always best to have the same language as everyone else in the population.

was introduced that regulates the probability of producing strings. When allowed to evolve, the parameter went quickly to 0, which meant that agents did not communicate at all anymore.

The paradox is thus that if you benefit from others understanding you, you shouldn't say anything novel, whereas if you don't benefit, it's best not to say anything at all. This paradox is a natural consequence of the problems of coordination and cooperation that I discussed in chapter 2, and it occurs also in simple models of the evolution of phonology and vocabulary. It is important, however, to work out solutions to this paradox that are relevant for the evolution of phrase-structure.

Pinker & Bloom (1990) suggest one such solution. They observe that for complex syntactic constructions, comprehension is always ahead of production, in acquisition, in use and perhaps also in evolution. We can thus imagine a scenario where syntactic innovations do offer an immediate selective advantage in the mutual benefit condition, because hearers can understand the new constructions even though they cannot actively use them themselves. It would be interesting to work out this proposal in a formal model.

Alternatively, under the hearer-benefit condition, a possibility might be that the willingness to speak is maintained through kin selection, whereas the evolution of complex syntax is driven by both the benefits of sharing more information with kin, and the benefits of making messages difficult to understand to non-kin. This could be called the **encryption hypothesis** (Will Lowe, p.c.). In Zuidema (2000) and Zuidema & Hogeweg (2000) we studied a simulation model where agents are placed on a 2d spatial grid. Preliminary results supported this hypothesis, but more formal work on this hypothesis, in the framework of social evolution theory (chapter 2), would be worthwhile. A related idea is what Fitch (2000) calls the "password hypothesis" about the evolution of complex patterns in bird song, which says that the function of these patterns is to distinguish kin from non-kin intruders.

In summary, there have been efforts by Fitch & Hauser (2004), Hashimoto & Ikegami (1996), myself and others to find a plausible scenario for the evolution of recursive, hierarchical phrase-structure using the classes of rewriting grammars from the Chomsky Hierarchy. In such a scenario, two or more stages are postulated (that is, finite-stateness and context-freeness), and experiments or simulations are performed to show that non-human primates are in one stage, and humans in another, or that simulated evolution can guide a population from one stage to the other.

These efforts run into a number of problems. First, it is clear that the actual strategy set available to evolution is much more constrained. For instance, the tamarins in the Fitch & Hauser

experiments do not *learn* $a^n b^m$ and the agents in the simulations cannot *parse* many of the finite-state and context-free languages that could have been useful. How to formalise the constraints on the strategy set from learning and parsing – that is, how to define the set of learnable languages $\mathcal{L}_A$ – is a difficult question. In simulations we can of course use specific learning, production and interpretation algorithms that implement these constraints without a formal characterisation of the set $\mathcal{L}_A$. This is the approach taken in this chapter, but it is somewhat unsatisfactory, because it is difficult to evaluate the generality of a specific simulation.

Second, it is not clear how to assign payoffs to the strategies in such a set. One proposal is to make payoff proportional to the number of strings in a language that are shared with others in the population. The problem is that this gives maximum potential fitness to grammars that accept any string over the alphabet used, which is clearly not what we see in natural language. Also, it seems that fitness in such a scheme is completely independent from the location on the Chomsky Hierarchy. More natural is a fitness scheme that is based not on whether a sentence can be parsed, but on whether it is interpreted correctly[10]. In simulation models we can explicitly incorporate a semantics – and reward correct interpretations – circumventing the problem of having to evaluate the payoff of grammars on a purely syntactic level. This is perhaps the best way forward, but again somewhat unsatisfactory because of the lack of generality.

Third, even if we choose a rather arbitrary strategy set and payoff function, it is still not trivial to show the evolution of phrase-structure, because of the problems of cooperation and coordination. Solutions to these problems remain an open issue.

6.2.3 *The Uniformity Assumption*

An approach that circumvents these problems and does give general results, is presented in Nowak, Komarova & Niyogi (2001) and follow-up papers (Komarova, Niyogi & Nowak, 2001; Mitchener & Nowak, 2002). This approach builds on a tradition of models of language known as Principles & Parameters (Chomsky, 1981), and of certain abstract models of learning in learnability theory (e.g. Bertolo, 2001; Niyogi, 1998). The core idea of the Principles & Parameters approach is to parameterise the variation in possible natural languages. That is, to find a description of natural languages where all differences between languages are characterised by different values of a set of binary parameters. With such a description in hand, the task of learning a language can now be described as the setting of these parameters. Theoretically, a description

[10]The assumed function of phrase-structure in models like Hashimoto & Ikegami (1996) and Zuidema & Hogeweg (2000) is simply to generate more messages, i.e. the quantitative advantages, and little attention has been paid to the likely advantages in the kind of semantic information that can be encoded and the kind of generalisations it allows in learning, i.e. the qualitative advantages. Perhaps the Chomsky Hierarchy is of more relevance there.

of language variation and acquisition that fits this scheme is a logical possibility, if the number of possible grammars is finite. In practise, however, such a description might be prohibitively complex.

The description simplifies enormously with the assumption that all possible grammars are equivalent in important ways. Such an "Uniformity Assumption" states, for instance, that all languages are equally useful for communication and equally easy or difficult to learn. Note that such an assumption implies that $\mathcal{L}_A = \mathcal{I}_A$, that is, it excludes the possibility that cultural evolution favours some languages over others. The mathematically convenient symmetry such an assumption introduces allows one to formulate simple models of language acquisition that use only the number of possible grammars and the number of available sentences to learn from as parameters. Note, however, that such an assumption forces one to treat grammatical rules as separate from the lexicon (as lexical variation clearly is unbounded), and to view child language development as a process of jumping from one adult grammar to another (because the Uniformity Assumption excludes growth in grammatical complexity). The assumption therefore goes hand in hand with a view that knowledge of language is specified in great detail in an innate Universal Grammar (UG)[11]. The term UG, in this tradition, can refer to both the innate language faculty and the set of possible grammars it allows.

In line with this tradition, Nowak *et al.* (2001) present a model of the evolution of Universal Grammar, where the UG is passed on to the next generation genetically and the actual grammars passed on culturally. Each individual "knows" one of the grammars from UG. That knowledge is passed on – strictly vertically – to her offspring, but mistakes are made in learning. If a mistake is made, the child ends up with a different grammar from the set of possible grammars. Finally, the authors assume that knowing a grammar i confers a fitness advantage that depends on the frequency of grammar i in the population. That is, speakers of the most frequent language receive most offspring. An evolving population, where the UG is constant, can thus be described in terms of the changes in the relative frequencies x_i of each grammar type i in the population, the probabilities Q_{ji} that a child ends up with grammar j when learning from her parent with grammar i, and the fitness F_i of a grammar i ($F_i = \sum_j x_j f_{ji}$, where f_{ji} gives the payoff from the communication between two individuals speaking i and j).

[11] Of course, many models of grammar learning have been proposed that do assume such an extensive UG, but do relax the Uniformity Assumption somewhat. For instance, Niyogi & Berwick (1995) assume a learning model where parameters are set based on triggers in the input data, which occur at different frequencies; Briscoe (2002a) incorporates the possibility of prior biases in his statistical parameter procedure; in Yang (2000) the language learner maintains multiple settings of parameters in parallel, and in learning converges to a single setting.

The first result that Nowak et al. obtain is a "coherence threshold". This threshold is the necessary condition for **grammatical coherence** in a population, that is, for a majority of individuals to use the same grammar. They show that this coherence depends on the chances that a child has to correctly acquire her parent's grammar. This probability is described with the parameter q. Nowak et al. show analytically that there is a minimum value for q to keep coherence in the population. If q is lower than this threshold value, all possible grammar types are equally frequent in the population and the average communicative success is minimal. If q is higher than this value, one grammar type is dominant; the communicative success is much higher than before and reaches 100% if $q = 1$.

The second result relates this required fidelity (q_1) to an upper and a lower bound on the number b of sample sentences that a child needs. The authors consider two learning strategies, that they claim represent the extremes on the possible strategies. The first is the "memoryless learner", that starts with a random grammar k in UG, and jumps to a random other grammar k' every time it is confronted with a training sentence that is inconsistent with grammar k. The second is the "batch learner" that memorises all b training sentences it receives, and picks a random grammar from the set of grammars consistent with all those sentences. The authors are aware, of course, that these learning algorithms are unrealistic. However, they argue that the algorithms represent an upper and a lower bound on the performance of realistic learning strategies.

In particular, the authors claim that the batch learner's performance is the best possible; the minimum number of sentences b_c it needs to reach the required fidelity q_1 for grammatical coherence therefore represents a lower bound for any learning algorithm. The authors show that b_c is proportional to the total number of possible grammars N. The actual number of sample sentences b is finite; Nowak et al. conclude that only if N is relatively small can a stable grammar emerge in a population. I.e. the population dynamics require a restrictive Universal Grammar (UG).

However, in a companion paper this claim is weakened. The authors write:

> *"Note that the number of candidate grammars can also be infinite, provided that children have a prior probability distribution specifying that some grammars are more likely than others". (Komarova, Niyogi & Nowak, 2001, p. 44)*

Here, they (correctly) present the finiteness of the hypothesis space as an assumption, rather than a conclusion from the model. The authors do, however, still make the claim that human language learning must have an intermediate performance in between that of the "memory-less learner" and the "batch learner". This claim is the core of the approach, because it allows the authors to study aspects of the cultural and biological evolution of language, without solving the problem

of specifying the set of languages that can be learnt by a specific algorithm ($\mathcal{L}_A$) or the set that results from cultural evolution ($\mathcal{I}_A$). In this chapter, however, I will show that this claim again depends on the uniformity assumption. In iterated learning, even a biased learning algorithm as mediocre as the one I present in section 6.3.1, will outperform the unbiased batch learner, because its bias is automatically the right one.

In the third part of Nowak *et al.* (2001), the authors consider the biological evolution of alternative UGs. In their scenario, a more restrictive UG can invade a population with a less restrictive UG, because it improves the learning accuracy. Mitchener & Nowak (2002) work out formally the conditions for such invasions, in simple examples with 1 or 2 UGs, that allow for 1 or 2 grammars. This analysis thus deals with the invasibility constraint, and provides a path of ever increasing fitness. However, the analysis gets extremely complex at times, even though only extremely simplified situations are considered. Moreover, it is unclear if the methods can be extended to deal with an non-uniform evolutionary scenario that starts with a limited proto-language and ends with a language approaching the complexity of modern language.

In short, Nowak and colleagues worked out an elegant framework to relate evolution and acquisition of grammars. If it could be adapted to deal with the evolution of phrase-structure per se, and other key features of syntax, that would be major progress. Unfortunately, the tractability of the equations depends largely on the simplifying assumption that all grammars are, in important senses, equivalent to each other. Such uniformity is a respectable null hypothesis in many issues in linguistics[12], but when applied to language evolution it is problematic (Newmeyer, 2003). Non-uniformity brings qualitatively different dynamics, as I will show below, but it is unclear how the framework can be extended to deal with such cases.

Implicit in the analysis, as in other learnability models (e.g. Gold, 1967; Wexler & Culicover, 1980), is the assumption that every possible grammar from some set is equally likely to become the target grammar for learning. If even the best possible learning algorithm cannot distinguish between all grammars from that set, the set of allowed grammars must be restricted. However, Expression–Induction models such as those discussed above give reason to believe that this assumption is not the most useful for language learning. Language learning is a very particular type of learning problem, because the outcome of the learning process at one generation is the input for the next. The samples from which a child learns with its learning procedure, are therefore *biased* by the learning of previous generations that used the same procedure (Kirby, 1994; Christiansen, 1994; Deacon, 1997). In the rest of this chapter I will develop a new computational model that

[12]Especially, of course, in political debates about the status of minority languages.

explores the consequences of this phenomenon for the thinking about the evolution of hierarchical phrase-structure.

6.3 Model Description

I have discussed a number of vastly different approaches to understand the origins of phrase-structure in natural languages. I first reviewed models known as Expression–Induction models, that show that from rather general learning procedures languages with a non-trivial recursive, hierarchical phrase-structure can emerge in a process of cultural evolution. I argued that these models fall short as *ultimate* explanations for the origins of phrase-structure, because they do not explain the origins of the particular learning abilities. The models do, however, have important implications, because they show that for non-trivial learning algorithms, the set of languages that can be represented $\mathcal{R}_A$, the set that can be learnt $\mathcal{L}_A$ and the set of languages that emerges in cultural evolution $\mathcal{I}_A$ are radically different sets.

I then reviewed two mathematically convenient ways to define a strategy-set, one based on the Chomsky Hierarchy and the other based on the Uniformity Assumption. It emerged from this review that in both approaches it is extremely difficult to incorporate the constraints from learning, parsing/production and cultural evolution. The classes of the Chomsky Hierarchy are not suitable to define a biological meaningful strategy set or payoff function. The Uniformity Assumption does allow one to define strategies and payoffs and to study invasibility, but the techniques to do so do not extend to systems without uniformity.

Hence, it seems the best way forward is to design and study simulation models – as Smith (2003b) did for learning strategies for vocabulary and simple compositionality – that do (i) explicitly incorporate computational procedures for the learning, interpretation and production of hierarchical phrase-structure, (ii) model a population with cultural transmission and cultural evolution, (iii) consider variants in these procedures, and (iv) associate different fitnesses with different outcomes. Such models are bound to become excessively complex, so we should search for minimal models that do include these components. In the following I will present the design of a model that represents a step in this direction. It will not yet incorporate all four of these components. In particular I will not deal with (iii), because I consider only a single learning algorithm. The goal is to design a model that includes components i, ii and iv, and is as simple as possible, but sufficiently rich to illustrate the problems with models based on the Chomsky Hierarchy and the Uniformity Assumption.

In the next section, 6.3.1, I will present a simple learning algorithm for context-free phrase-structure grammars. In section 6.4.1 I briefly discuss its learning abilities, and in section 6.4.2

I will present a simple iterated learning model. Finally, in section 6.4.3 I include natural selection, and show how cultural evolution makes the dynamics deviate from predictions in existing work.The model I develop in these sections thus integrates learning, cultural evolution and natural selection. Although it has so far little to say about the evolutionary origins of the ability to learn phrase-structure per se, I hope it represents another step toward an formal account of the evolution of phrase-structure.

6.3.1 A Simple Model of Grammar Induction

The first step is to design a grammar induction algorithm that is simple, but can nevertheless deal with some non-trivial induction problems. The algorithm uses context-free grammars to represent linguistic abilities. In particular, the representation is limited to grammars where all rules are of one of the following forms: (1) $A \mapsto t$, (2) $A \mapsto BC$, (3) $A \mapsto Bt$. The nonterminals A, B, C are elements of the non-terminal alphabet V_{nt}, which includes the start symbol S. t is a string of terminal symbols from the terminal alphabet V_t.

Hence, we can easily characterise the set of language representable by this algorithm ($\mathcal{R}_A$). Note that, beyond context-freeness, the restrictions on the rule-types above do not limit the scope of languages that can be represented: rule types (1) and (2) are those of the Chomsky Normal Form, with which, as is well known, any context-free language can be modelled. They are, however, relevant for the language acquisition algorithm; rule type (3), for instance, allows a simple formulation of the compression step (described below), such that only a single new non-terminal needs to be introduced at every learning step. Note further that the class of languages that this formalism can represent is unlearnable by Gold's criterion (Gold, 1967). That is, there will always be multiple hypotheses consistent with the training data, such that the target grammar can not be uniquely identified. Note finally that the model involves no semantics. Although I believe semantics plays a major role in language learning and language evolution, as explored also in chapter 5, the goal of this chapter is to evaluate the applicability of the Chomsky Hierarchy and the Uniformity Assumption in these questions. Semantics has so-far not played much of a role in the discussion; my approach is therefore to see how far we get without the extra complication.

For determining the language L of a certain grammar G I use simple depth-first exhaustive search of the derivation tree. For computational reasons, the depth of the search is limited to a certain depth d, and the string length is limited to length l. The set of sentences ($L' \subseteq L$) used in training and in communication is therefore finite. In production, strings are drawn from a uniform distribution over L'. In the communication between two agents, the speaker chooses a random element s from her language, and the hearer checks if s is an element of his own language. If so,

the interaction is a success, otherwise it is a failure (the success of interactions will play a role in the version of model that includes natural selection, as will be discussed below).

The language L is generated by calling the function `search-subtree(`"S"$,d)$, which can be defined in pseudo-code as follows:

```
% l is the maximum string length parameter.
% L is an initially empty set.
search-subtree(s, d')
   if (d' < 1 OR LENGTH(s) > l) stop
   if (ALLTERMINAL(s)) add s to language L
   for all rules r
      for each fit of r on s
         apply r to s, yielding s'
         search-subtree(s', d' − 1)
```

The grammar induction algorithm used in the model consists of three operations: (i) incorporation, (ii) compression and (iii) generalisation. The learner learns from a set of sample strings (sentences) that are provided by a teacher. The design of the learning algorithm is originally inspired by Kirby (2000) and is similar to the algorithm in Wolff (1982). The algorithm fits within a tradition of unsupervised grammar induction algorithms that search for compact descriptions of the input data (e.g. Solomonoff, 1960; Stolcke, 1994; Rissanen & Ristad, 1994; van Zaanen & Adriaans, 2001). The three operations are defined as follows:

Incorporation: *extend the language, such that it includes the encountered string*; if string s is not already part of the language, add a rule $S \mapsto s$ to the grammar.

Compression: *replace frequent and long substrings with a nonterminal, such that the grammar becomes smaller and the language remains unchanged*; for every valid substring z of the right-hand sides of all rules, calculate the compression effect $v(z)$ of replacing z with a nonterminal A; replace all valid occurrences of the substring $z' = argmax_z v(z)$ with A if $v(z') > 0$, and add a rule $A \mapsto z'$ to the grammar. "Valid substrings" are those substrings which can be replaced while keeping all rules of the forms 1–3 described above. The compression effect is measured as the difference between the number of terminal and non-terminal symbols in the grammar before and after the replacement (i.e. the sum length of all rules in the grammar). The compression step is repeated until the grammar does not change anymore. At every step, the number of non-terminal symbols increases by 1.

Generalisation: *equate two nonterminals, such that the grammar becomes smaller and the language larger*; for every combination of two nonterminals A and B ($B \neq S$), calculate the compression effect v of equating A and B. Equate the combination $(A', B') = argmax_{A,B} v(A, B)$ if $v(A', B') > 0$; i.e. replace all occurrences of B with A. The compression effect is again measured as the difference between the number of symbols before and after replacing and deleting redundant rules. The generalisation step is repeated until the grammar does not change anymore. At every step, the number of non-terminal symbols decreases by 1.

For the grammar acquisition algorithm these three operations can be used in several setups. For the purposes of this chapter, I have chosen simple *off-line learning*: compression and generalisation occur after all training strings have been received. I have added one additional step to this basic algorithm. To avoid insufficient expressiveness, the generalisation phase concludes with a check for the size of the language. If this size is smaller than some minimum, $E = size(L) < E_M$, I generate $E_M - E$ random strings[13] and `incorporate` them in the grammar. This procedure can be considered a substitution for the semantics that is left out in the model, because it prevents the simulation to collapse to the perfectly learnable, but totally pointless language with just a single sentence (expressiveness $E = 1$). Thus, in pseudo-code, the learning algorithm is:

```
% i,j are agent objects, each with their own grammar G and language L;
% l_0 is a parameter for the initial string length;
% V_te is the terminal alphabet;
teach(i,j)
   repeat T times
      i generates random string s from L_i
      j calls incorporate(s)
   repeat until G_j does not change anymore
      j calls compress()
   repeat until G_j does not change anymore
      j calls generalise()
   if (E < E_M)
      repeat E_M - E times
         generate a random string s ∈ (V_te)* of size ≤ l_0
```

[13] These strings have a maximum size l_0 which is an important parameter in the results section ($l_0 \leq l$).

```
j calls incorporate(s)
```

6.4 Results

6.4.1 *Learnable and Unlearnable Grammars*

The algorithm described above is implemented in C^{++} and tested on a variety of target grammars. I will not present a detailed analysis of the learning behaviour here, but limit myself to a simple example that shows that the algorithm can learn some (recursive) grammars, while it cannot learn others. The induction algorithm receives three sentences (abcd, abcabcd, abcabcabcd). The incorporation, compression (repeated twice) and generalisation steps (without the extend expressiveness step, i.e. with $E_M = 0$) yield subsequently the following grammars:

(a) Incorporation	(b) Compression	(c) Compression	(d) Generalisation
S $\mapsto$ abcd	S $\mapsto$ abcd	S $\mapsto$ Yd	S $\mapsto$ Xd
S $\mapsto$ abcabcd	S $\mapsto$ Xd	S $\mapsto$ Xd	S $\mapsto$ Xabcd
S $\mapsto$ abcabcabcd	S $\mapsto$ Xabcd	S $\mapsto$ Xabcd	X $\mapsto$ XX
	X $\mapsto$ abcabc	X $\mapsto$ YY	X $\mapsto$ abc
		Y $\mapsto$ abc	

In (b) and (c) the substrings "abcabc" and "abc" are subsequently replaced by the non-terminals X and Y. In (d) the non-terminals X and Y are equated, which leads to the deletion of the second rule in (c). One can check that the total size of the grammar reduces from 24, to 21 and further down to 19 and finally 16 characters.

From this example it is also clear that learning is not always successful. Any of the four grammars above ((a), (b) and (c) are weakly equivalent, i.e. generate the same string language) could have generated the training data, but with these three input strings the algorithm yields grammar (d). Many target grammars will never be learnt correctly, no matter how many input strings are generated. In practise, each finite set of randomly generated strings from some target grammar, might yield a different result. Thus, for some number of input strings T, some set of target grammars are always acquired, some are never acquired, and some are some of the time acquired. This variation in difficulty is an important precondition for the occurrence of cultural evolution.

If we can enumerate all possible grammars, we can describe this with a matrix $\mathbf{Q}$, where each entry $\mathbf{Q}_{ij}$ describes the probability that the algorithm learning from sample strings from a target grammar i, will end up with grammar j (since the learning algorithm is deterministic, this means that $\mathbf{Q}_{ij}$ gives the proportion of possible sets of example sentences – "texts" in learnability

theory jargon – generated by a parent grammar i that will lead to a child grammar j). $\mathbf{Q}_{ii}$ is the probability that the algorithm finds target grammar i.

We can now be a bit more precise about "learnability" and "learnable languages". A class of languages $\mathcal{L}$ is learnable, in Gold's (1967) sense of "identification in the limit", only if there exists an algorithm that can learn all languages i in that class in the limit of infinitely many example sentences: $\exists A \forall i \left[\lim_{T\to\infty}(\mathbf{Q}_{ii}^{A,T} = 1)\right]$. For grammatical coherence, in the sense of Nowak *et al.* (2001), we need for a specific algorithm A, a specific dominant language i and a given number of training samples T, the $\mathbf{Q}_{ii}^{A,T}$ to be above a threshold value q_1.

A different definition of learnability is based on the degree of similarity (or expected communicative success) between a target grammar i and the learnt grammar j, where j is induced with algorithm A from T training samples generated by i. This value $\mathbf{C}_{ij}^{A,T}$ can be estimated by counting how many out of a finite sample of strings generated by grammar i are accepted by grammar j. In the idealised model of Nowak *et al.* (2001) all grammars/languages are at equal distance a from each other, and hence $\mathbf{C}_{ij}^{A,T} = a$ if $i \neq j$, and $\mathbf{C}_{ij}^{A,T} = 1$ if $i = j$. In such a model, only the probability that the correct grammar is induced is relevant. In the current model, however, the probability that the induced grammar is identical to the target grammar is vanishingly small for all interesting grammars. More relevant is whether or not the induced grammar is sufficiently similar to the target grammar. This value will be reported in the simulation results.

Given the various concepts of learnability, how should we define the set of learnable languages $\mathcal{L}_A$? In the current model it is probably best to think about this set as those languages with a C-score above a given base-line. Given the limitations of the learning algorithm, how do we ensure that learning is successful? The following section will show that for this we need nothing more than to assume that the output of one learner is the input for the next.

6.4.2 *Iterated Learning: the Emergence of Learnability*

To study the effects of iterated learning, I extend the model with a population structure. In the new version of the model individuals (agents, that each represent a generation) are placed in a *chain*. The first agent induces its grammar from a number E of randomly generated strings. Every subsequent agent (the child) learns its grammar from T sample sentences that are generated by the previous one (the parent). Using the matrix $\mathbf{Q}$ from the previous section, we can formalise this *iterated learning model* with the following general equation, where x_i is the probability that

grammar i is the grammar of the current generation:

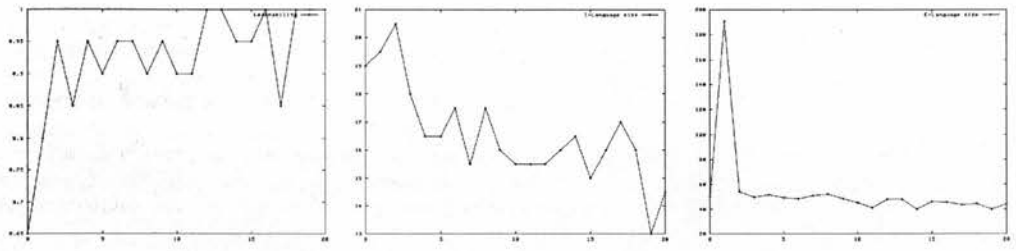
$$\Delta x_i = \sum_{j=0}^{N} x_j \mathbf{Q}_{ji} \tag{6.3}$$

In simulations such as the one of figure 6.5 the communicative success $\mathbf{C}_{ij}$ between child j and parent i rises steadily from a low value (here 0.65) to a high value (here 1.0). In the initial stage the grammar shows no structure, and consequently almost every string that the grammar produces is idiosyncratic. A child in this stage typically hears strings like "ada", "ddac", "adba", "bcbd", or "cdca" from its parent. It cannot discover many regularities in these strings. The child therefore cannot do much better than simply reproduce the strings it heard (i.e. T random draws from at least E_M different strings; with the given parameters, the expected C-score is 0.65, which is the baseline, above which I will call a language learnable), and generate random new strings if necessary, to make sure its language obeys the minimum number (E_M) of strings.

However, in these randomly generated strings, sometimes regularities appear. I.e., a parent may use the randomly generated strings "dcac", "bcac", "caac" and "daac". When this happens the child tends to analyse these strings as different combinations with the building block "ac". Thus, typically, the learning algorithm generates a grammar with the rules $S \mapsto dcX$, $S \mapsto bcX$, $S \mapsto caX$, $S \mapsto daX$, and $X \mapsto ac$. When this happens to another set of strings as well, say with a new rule $Y \mapsto b$, the generalisation procedure can decide to equate the non-terminals X and Y. The resulting grammar can then generalise from the observed strings, to the unobserved strings "dcb", "bcb", "cab" and "dab". The child still needs to generate random new strings to reach the minimum E, but fewer than in the case considered above.

Now consider the next step in the simulation, when the child becomes itself the parent of a new child. This child is presented with a language with more regularities than before, and has a fair chance of *correctly* generalising to unseen examples. If, for instance, it only sees the strings "dcac", "bcac", "caac", "bcb", "cab" and "dab", it can, through the same procedure as above, infer that "daac" and "dcb" are also part of the target language. This means that (i) the child shares more strings with its parent than just the ones it observes and consequently shows a higher between generation communicative success, and (ii) regularities that appear in the language by chance, have a fair chance to remain in the language. In the process of iterated learning, languages can thus become more structured and better learnable.

Similar results with different formalisms were already reported before (e.g. Kirby, 2000; Brighton, 2002), but here I have used a standard grammar formalism (context-free grammars)



(a) Learnability $\mathbf{C}_{ij}$ (b) Number of rules (c) Expressiveness

Figure 6.5: Iterated Learning: although initially the target language is unstructured and difficult to learn, over the course of 20 generations (a) the learnability ($\mathbf{C}_{ij}$, where i is the grammar at time $t-1$ and j the grammar at time t.) steadily increases, (b) the number of rules steadily decreases (combinatorial and recursive strategies are used), and (c) after a initial phase of over-generalisation, the expressiveness remains close to its minimally required level. Parameters: $V_t = \{a, b, c, d\}$, $V_{nt} = \{S, X, Y, Z, A, B, C\}$, T=30, E=20, l_0=3. Shown are the average values of 2 simulations.

without semantics and an extremely simple learning algorithm. The results show that the effects of iterated learning do not depend on semantics or other idiosyncratic features of previous models. Furthermore, the model provides the simplest illustration of the fact that in iterated learning representable but unlearnable languages will disappear, and that over the course of a number of generations the languages can become better and better learnable.

It is interesting to contrast the observations here with popular interpretations (e.g. Wexler, 1999; Bertolo, 2001) of negative results in learnability theory, such as Gold's proof (Gold, 1967). Whereas in the usual interpretations of that proof it is assumed that we need innate constraints on the *search space* in addition to a smart *learning procedure*, here I show that even a simple learning procedure can lead to successful acquisition, because restrictions on the search space automatically emerge in the iteration of learning. If one considers learnability a *binary* feature – as is common in for instance Principles & Parameters theory – this is a rather trivial phenomenon: languages that are not learnable will not occur in the next generation. However, if there are gradations in learnability, the cultural evolution of language can be an intricate process where languages get shaped over many generations.

6.4.3 Language Adaptation and the Coherence Threshold

When we study this effect in a version of the model where *selection* does play a role, it is also relevant for the evaluating the claims of Nowak *et al.* (2001) and Komarova *et al.* (2001). The model is therefore extended such that every generation consists of P agents. We can now not only measure the success in communicating with a parent (the between generation success, $C_{between}$),

but also the success in communicating with other agents of the same generation (the within generation success, C_{within}). Hence, we have at every point a population of parents and a population of children. Children communicate with each other; when they become parents themselves the expected number of offspring of any one of them (the *fitness*) is determined by the number of successful interactions it had. Children still acquire their grammar from sample strings produced by their parent. Adapting equation 1, this system can now be described with the following equation (where, assuming an infinite population size, x_i is now the relative fraction of grammar i in the population):

$$\Delta x_i = \sum_{j=0}^{N} x_j f_j \mathbf{Q}_{ji} - \phi x_i \tag{6.4}$$

Here, f_i is the *relative fitness* (quality) of grammars of type i and equals $f_i = \sum_j x_j F_{ij}$, where F_{ij} is the expected communicative success from an interaction between an individual of type i and an individual of type j. The relative fitness f of a grammar thus depends on the frequencies of all grammar types, hence it is *frequency dependent.* ϕ is the average fitness in the population and equals $\phi = \sum_i x_i f_i$. This term is needed to keep the sum of all fractions at 1.

This equation is essentially the model of Nowak *et al.* (2001). Recall that the main result of that paper is a "coherence threshold": there is a minimum value q_1 for the learning accuracy q to keep coherence in the population. If $q < q_1$, then coherence is lost because natural selection is not effective; if $q > q_1$, one dominant grammar emerges. We can try to apply the calculations of Nowak et al. to a simulation with the learning algorithm and selection as described above. In the simulation I will report below, learners are presented $T = 100$ sample sentences, with initial string length $l_0 = 12$ and an alphabet of 4 characters.

The model of Nowak et al. uses three parameters: a (the similarity between two different languages), T and N (the number of possible languages). The applicable expression for q_1 is the following:

$$q_1 = \frac{2\sqrt{a}}{1+\sqrt{a}}. \tag{6.5}$$

Assuming a string length $l = 12$, we can make an estimate of a as follows. There are 4^{12} strings of length 12 over the given alphabet. Assuming languages with 100 strings, the probability that a random string s is part of the language L is $P(s \in L) = \frac{100}{4^{12}}$. Given the large number of strings, the similarity a between two random languages can be estimated as $a \approx 100 \cdot P(s \in L) = \frac{10000}{4^{12}}$. Plugging this value in equation (6.5) yields an estimate of the "coherence threshold" at learning accuracy $q_1 \approx 0.05$.

Crucial for grammatical coherence is that the accuracy of a given learning algorithm is above this value. Nowak et al.'s calculation of the accuracy of the batch learner, q_{batch}, is as follows:

$$q_{batch} = \frac{\left(1 - \left(1 - a^T\right)^N\right)}{(Na^T)}. \tag{6.6}$$

Assuming string length and number of strings as above, an estimate of N is:

$$N \approx \begin{pmatrix} 4^{12} \\ 100 \end{pmatrix} \approx 10^{520}. \tag{6.7}$$

Hence, Na^T, the divider in equation (6.6), will be astronomical, on the order of 10^{420}. Because the value of the denominator will be between 0 and 1, it follows that $q_{batch} \approx 0$ and therefore $q_{batch} << q_1$. That is, the batch learner will, with so many possible languages and so few training samples, have almost no chance of discovering the target grammar, and the system should therefore be below the coherence threshold.

Below the threshold, all languages occur at equal frequency, and we therefore expect the within generation communicative success to be equal to the similarity between languages, i.e. $C^*_{within} = a \approx 0.05$. This result depends on a number of strong assumptions. For instance, Nowak *et al.* (2001) assume that the population is infinite, and that all grammars are of equal quality (uniformity), such that only the frequency of a grammar in the population determines the payoff it will receive. But most importantly, the authors assume that all possible grammars are at equal distance a from each other, whereas in the present simulation model many distances are possible. These assumptions are all violated in the present model, as well as in reality of course[14]. Before discussing how to adapt the model of Nowak et al., I will first present the results from the simulation model and compare them with the calculations above.

Figure 6.6 shows results from a simulation with the grammar induction algorithm described earlier in this chapter. Unlike the simulations of figure 6.5, these simulations deal with a relatively difficult learning problem: here the initial string length is long, i.e. $l_0 = 12$, whereas before it was $l_0 = 6$. Learning is therefore not successful. In the left region of the graph it can be seen that the between-generation C is around 70%, which (with the given parameters) means that only the strings a child has seen during training, are shared between a parent and a child. The child has not been able to make any correct generalisations. Qualitatively, we see that the simulation in this

[14]See Wiehe (1997) for a critique of the error threshold model, on which the analysis of Nowak *et al.* (2001) is based.

phase is indeed below the coherence threshold. There is no dominant grammar in the population, and agents score only about 15% in their communication with peers. This is reasonably close to the predicted 5%; the disparity is explained by mentioned unrealistic assumptions[15].

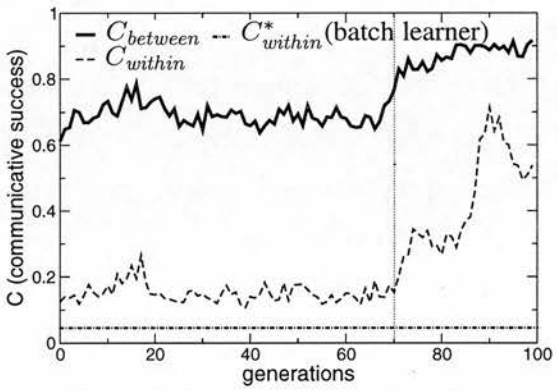


Figure 6.6: Results from a run under fitness proportional selection. This figure shows that there are regions of grammar space where the dynamics are apparently under the "coherence threshold" (Nowak *et al.*, 2001), while there are other regions where the dynamics are above this threshold. The parameters, including the number of sample sentences T, are still the same, but the language has adapted itself to the **bias** of the learning algorithm. Parameters are: $V_t = \{0, 1, 2, 3\}$, $V_{nt} = \{S, a, b, c, d, e, f\}$, P=20, T=100, E=100, l_0=12. Shown are the average values of 20 agents.

However, around generation 70 the behaviour of the simulation starts to diverge radically from these analytical predictions. First, the between generation communicative success suddenly rises. Children are now able to successfully generalise beyond the strings that are seen during training, and score around 90% in communicating with their parents. The reason is the same as in the previous section: the languages have adapted to the learning algorithm. Crucially, the grammatical coherence in the population also rises. In this second phase, agents score between 30% and 70% in the communication with peers. With always the same T (number of sample sentences), and with always the same grammar space, there are regions where the dynamics are apparently under the "coherence threshold", while there are other regions where the dynamics are above this threshold. The language has adapted to the learning algorithm, and, consequently, the coherence in the population does not satisfy the prediction of Nowak et al.

Recall that Nowak, Komarova & Niyogi (2001) and Komarova, Niyogi & Nowak (2001) considered an upper and a lower bound on the performance of learning algorithms, and claimed that the performance of human language learning must be in between those bounds. That is, they claimed that (i) the batch learner represents the best possible learning strategy, and that (ii) the number of training samples it requires for grammatical coherence, given by equations (6.5)

[15] Because of the variation in distances between languages, we can expect the "effective similarity" to be higher, and therefore the actual value of q_1 and the equilibrium C_{within} as well.

and (6.6), represents a lower bound on the number of samples children require, and hence a fundamental requirement for effective biological evolution of the language faculty. Figure 6.6 falsifies that the claim. It shows that in iterated learning, a biased algorithm can do better than the unbiased batch learner, because over time the languages will adapt to the bias.

Although it is clear that the simulation of figure 6.6 is inconsistent with the calculations in Nowak *et al.* (2001) and Komarova *et al.* (2001), my results do not yet unambiguously prove that the batch learner, if implemented with the same parameters as in the current simulation, would lead to worse grammatical coherence in the population. The difficulties I see with demonstrating a difference in performance for this particular case are many. First, it is extremely difficult to implement the batch learner in a computer program, because of the astronomical number of possible languages (although I cannot exclude the possibility that an ingenious way exists to encode which subset of 10^{520} possible languages are consistent with the training data, and to choose a random language from a uniform distribution over that subset). Second, although the equations above could in principle be adapted to include the effects of unequal distances, the math quickly gets very complicated. The problem is that equations (6.5) and (6.6) are not valid if a is an average rather than a constant.

On a more general level, however, there is no doubt that the central claim of Nowak *et al.* (2001) and Komarova *et al.* (2001) that the batch learner provides a lower bound on the number of training samples needed in human language acquisition is untenable. The batch learner is an unbiased learner: it chooses a random language from all those possible languages that are consistent with the T sample sentences received. It is unbeatable only if all possible languages have equal probability of becoming the target language. If, however, there is a skew in the distribution of possible target languages, and if a learning algorithm is biased towards the more likely targets, that algorithm can beat the unbiased learner. The iterated learning model studied in this chapter provides the mechanism by which this situation will arise in culturally transmitted systems such as language.

6.5 Conclusions

I believe that these results have some implications for our thinking about both language acquisition and language evolution. In particular, I think the model and results offer a different perspective on the argument from the poverty of the stimulus, and thus on one of the most central "problems" of language acquisition research: *the logical problem of language acquisition.* This is the problem every child is facing when acquiring the grammar of its native language: she has

insufficient evidence to uniquely determine which is the target grammar from the set of all grammars that would in principle be possible. I will discuss the new perspective evolutionary models offer in the wider debate about such arguments in the next chapter.

Formal models of language learning often specify three components (Bertolo, 2001). The first is the innate knowledge of language a child has, which defines a hypothesis space for language learning. The second is the primary linguistic data, which provides the input to the learning procedure. The third is the learning procedure, that guides the child through the hypothesis space based on the input data it encounters. Using such a scheme, one might be tempted to describe the model I presented in this chapter as follows: it uses context-free grammars as the hypothesis space, unordered positive examples ("text") as primary linguistic data and a simple greedy, compression based heuristic as the search procedure. How can learning be successful, given the often cited mathematical proof (Gold, 1967) that learning under these conditions is impossible?

The answer is of course that the question is wrong. Gold's proof is about certain classes of formal languages and establishes that no algorithm can be *guaranteed* to learn any language from such a class. It makes no claim about whether or not a specific algorithm will be able to learn lots of specific languages from the class. Gold's proof only implies that an algorithm cannot learn *all* the languages from the class. Accepting for a moment Gold's definition of learnability (which is different from the definition I adopted), the proof shows that not all context-free languages can naturally occur as culturally transmitted codes.

Now, what determines which languages do, and which ones don't? The present model illustrates that it is unnecessary to assume – as theorists of language acquisition have often done (Wexler & Culicover, 1980; Bertolo, 2001) – additional, innate knowledge of language to constrain the set of naturally occurring languages. The only languages a child will ever need to learn, are languages learnt and transmitted by previous generations. Hence, the poverty of the stimulus is never a problem; rather, the ancestors' poverty is the solution to the child's.

The relevance of this perspective for language evolution is two-fold. First, Gold's proof and other mathematical learnability results are often – erroneously – cited as proof for domain-specific, innate knowledge of language (Wexler, 1999), of a sort that could only have arisen if it were selected for in human evolution (Pinker & Bloom, 1990). It is important to note that although the conclusion may be correct, the arguments for it are not (Pullum & Scholz, 2002; Scholz & Pullum, 2002; Johnson, 2004). Parsimony suggests that the natural *explanandum* for language evolution research is the learning strategy (which, in a process of cultural evolution, gives rise to languages with the structure we observe today), and certainly not an ornate, innate

database of syntactic rules and principles, as some naive theories of language acquisition have it (e.g. Cook, 1993).

Second, a methodological point for evolutionary linguistics is that the definition of a strategy set and a payoff function must take into account the constraints from learning, usage and cultural transmission. This might seem an obvious point, but it is an extremely difficult task. It is therefore tempting to simplify the description of learning by considering random walk or parameter setting models as upper and lower bounds on the performance of the *real* learning strategy. The problem with this approach is that – through cultural evolution – the *learning strategy* will change the *learning task*. Derived lower and upper bounds might therefore no longer be valid.

The model I presented does suggest some promising future directions for establishing more positive results on the evolution of phrase-structure. The model incorporates most of the components of an evolutionary analysis – only variation in the learning strategies is missing. A simple approach to introduce variation would be to vary the parameters of the model: the terminal and non-terminal alphabet, the initial string length l_0, the minimal expressiveness E_M and the number of training samples T. I haven't implemented such an extended model only because I believe real progress on these issues will come from a second future direction: the development of better, more robust and more flexible learning algorithms.

CHAPTER 7

Conclusions

7.1 Summary

I started this thesis, in **chapter 1**, by arguing that theories of the evolutionary origins of language need to be formalised and to entail complete scenarios of the transitions necessary to get from primate-like communication to modern human language. Theories with these features can be tested both by confronting them with empirical data, and by evaluating their internal coherence and consistency with other tested theories. In this thesis I have tried to contribute to such a testable theory, by surveying the formal requirements on such scenarios (drawing on insights from mathematical population genetics, evolutionary game theory and theoretical linguistics), and by proposing specific models for some of the transitions involved.

In **chapter 2** I reviewed formal, foundational models from evolutionary biology. This review yielded a list of criteria for evolutionary scenarios. The evolutionary process is described by (1) heritable traits, (2) a strategy set and (3) a payoff function. To show that a strategy can evolve, one needs to show that it can (4) invade a population without it, and that (5) there is a path of fit intermediates from the hypothesised start point to the end point of evolution. Furthermore, one needs to show (6) that the mutational load has not been too high, and that (7) the predicted time to fixation of innovations is realistic. These requirements lead to two main problems in the evolution of communication and language, which I termed (8) the problem of cooperation, and (9) the problem of coordination, respectively. Finally, I argued that evolutionary explanations need to (10) specify and relate the assumed levels of selection and heritability.

In **chapter 3** I reviewed a specific scenario for the evolution of language, proposed by Ray Jackendoff (2002). Jackendoff lists a number of different stages, and hints at the selective advantages of each of the innovations. However, his treatment remains very informal. Based on an

inventory of formal models from linguistics, I argued that there are three "major transitions" in the scenario that can be formally characterised: the emergence of combinatorial phonology, of compositional semantics and of hierarchical phrase-structure. An important distinction in such characterisations is that between I-language, the language system internal to the language user, and E-language, the observable language external to the user.

In **chapter 4** I studied a model that addressed the first of these transitions: the evolution of combinatorial phonology. After arguing that existing models fail some of the listed requirements, I considered a new model where the strategy set consists of all configurations of a set of trajectories in an acoustic space. I then proposed a measure of confusability as the payoff function. In simulations, I showed that combinatoriality in the E-language can emerge, without the I-language necessarily having that property. I argued that in a population where the E-language is combinatorial, an I-language which takes advantage of that fact can more easily evolve.

In **chapter 5** I discussed the evolution of compositional semantics. The strategy set of the model I presented, included all possible $\mathbf{S}$ (production) and $\mathbf{R}$ (interpretation) matrices, which represent a mapping from a meaning space to a signal space and vice versa. I considered a payoff function where the payoff depends on the probability of correctly interpreting a received signal. This function took into account noise and coordination. Crucially, I considered the case where, if the interpretation was different from the intended meaning, the payoff was still higher for interpretations more similar to the intention than for those less similar. In simulations, I showed that with these assumptions the mapping between sounds and meanings can become structured (superficial compositionality), without the underlying cognitive procedures necessarily making use of that property (productive compositionality). Analogous to chapter 4, I argued that in a population with superficial compositionality, productive compositionality can more easily evolve.

In **chapter 6** I focused on the evolution of recursive, hierarchical phrase-structure. Formulating a reasonable strategy set and payoff function for phrase-structure is difficult. I reviewed a number of existing models, and then presented a new computational model that integrated many of their features. From this study it emerged that, in general, the set of languages $\mathcal{R}_A$ that can be represented, the set $\mathcal{L}_A$ that can be learned, and the set $\mathcal{I}_A$ of languages that are stable in cultural transmission, are all different. Moreover, the relation between $\mathcal{R}_A$, $\mathcal{L}_A$, $\mathcal{I}_A$ and the learning strategy A can be extremely complex. Definitions of a strategy set based on $\mathcal{R}_A$, such as those based on the Chomsky Hierarchy (which ignore learning), and definitions based on the Uniformity Assumption (which exclude cultural evolution and imply that $\mathcal{R}_A = \mathcal{L}_A = \mathcal{I}_A$), are therefore unsatisfactory. The computational model I developed in this chapter showed the feasibility of a model

that integrates learning, cultural evolution and natural selection. The results from the simulations have important consequences for theorising about the acquisition and evolution of syntax. They do not, however, present a satisfactory explanation for the third major transition – the evolution of hierarchical phrase-structure – because the model did not yet present a reasonable set of possible learning strategies. This issue is left for future work.

7.2 Contributions

The primary goal of this thesis has been to contribute to the development of scientifically rigorous theories of the origins of human language. Formalisation and empirical testability are generally seen as the key features of scientific theories. Although not of the level of mathematical and empirical sophistication of some subfields of linguistics and biology, I hope this thesis includes some contributions to this end in a field where consensus on the goals and requirements of research is still lacking. Another important criterion for scientific theories is the potential to connect them with theories in related fields. I believe there are many such connections to be made, both within linguistics and within biology. In this section I will discuss some possible contributions of this thesis concerning testability; in sections hereafter I will raise some implications for related fields.

In this thesis, I have adopted a gradual scenario, with the complexity of modern language evolving in a number of steps rather than in one "big bang". With Pinker & Bloom (1990), Jackendoff (2002) and many others, I have assumed that the driving force behind this evolution has been the need to reliably convey more and more information. In chapters 2 and 6 I briefly mentioned alternative selection pressures where complex language serves to impress peers or sexual partners (e.g. Dessalles, 1998), or to communicate with kin (e.g. Fitch, 2004). Of course, it is impossible to measure empirically which of these selection pressures really was responsible for each of the transitions. However, we can evaluate the coherence of a complete scenario if it is precise and formal enough, and we can derive specific predictions from such a scenario that can be empirically tested.

This thesis has presented a framework, rather than a complete scenario. This was most obvious in chapter 6, where a reasonable strategy set was still lacking. But also in chapter 4 and 5 I have not committed myself to a very specific scenario. In these chapters I argued that the hillclimbing heuristic used can be interpreted as describing either evolutionary optimisation, or optimisation through learning, or a combination of both[1]. Nevertheless, I think in either interpre-

[1]The validity of this argument can be further investigated in simulation models where the processes of evolution and learning are modelled at a more concrete level.

tation, there are many starting points for empirical research. Some examples of specific testable issues are:

- In chapter 4 I proposed the use of *trajectories* through a low-dimensional acoustic space as the representation of signals with non-zero duration. Although I provided some examples of real acoustic data using this representation (courtesy of Bart de Boer), it remains an empirical issue whether signals in animal and human communication can in general be adequately described in this representation.
- Also in chapter 4, I presented a measure for the quality of a repertoire of signals, based on the average probability to confuse those signals. The function to relate distance to confusion probabilities was motivated by an extremely idealised case of just two signals in a one-dimensional acoustic space with Gaussian noise. The shape of the distance-to-confusion function in more realistic circumstances can be empirically estimated.
- The model of chapter 4 predicts that combinatorial repertoires are less confusable than non-combinatorial systems within the same acoustic and time constraints. This prediction can be tested empirically, both in humans and in other animals.
- More theoretical work is needed on a measure for how combinatorial a signal repertoire is, but within the same conceptual framework this could in principle be measured in human languages and animal communication systems. This would yield comparative data on how unique combinatorial phonology really is in nature.
- In chapter 5 I used a measure for the degree of topology preservation (proposed by Brighton, 2003). If acoustic and semantic similarity are operationalised, this measure could be applied to natural language data (as worked out by Shillcock *et al.*, 2004). It would be interesting to apply this measure to empirical data on primate communication as well.
- The model in chapter 6 suggests languages adapt to the language users in order to become easier to learn. It would be interesting to analyse empirical data on creolisation (Bickerton, 1990) and rapid language change in Nicaraguan sign language (Senghas *et al.*, 2004) from this perspective.

Possible contributions toward formalisation, and thence testability of the internal coherence of theories of language evolution are:

- The list of formal requirements on scenarios of the evolution of language from the perspective of evolutionary biology in chapter 2.
- The steps toward formalisation of Jackendoff's scenario for the evolution of language in chapter 3.

- The strategy set using the trajectory representation in chapter 4, which is an improvement over representations used in Nowak & Krakauer (1999) and Liljencrants & Lindblom (1972), but very similar to the representation used in Oudeyer (2002). The relation with trajectory models in speech recognition remains to be explored (Goldenthal, 1994).
- The payoff function based on confusability in chapter 4, which is an improvement over the measures used in Liljencrants & Lindblom (1972); Lindblom *et al.* (1984); the relation with work in phonetics and signal detection theory remains to be explored.
- The payoff function that includes the "value matrix" $\mathbf{V}$ in chapter 5, which is an improvement over the measures used in for instance, Oliphant & Batali (1996), Steels (1995) and Nowak & Krakauer (1999); the relation with work in distortion theory and compression theory remain to be explored.
- The learning algorithm in chapter 6, which is a simplification of algorithms presented in, for instance, Wolff (1982), Stolcke (1994) and Kirby (2000).
- The distinction between the sets of representable, learnable and stable languages in chapter 6.

7.3 Implications for Linguistics

I believe the main implication of this and related work for the field of linguistics concerns a debate we can call the "nativism–empiricism debate" (even though these terms are overstating the positions of actual researchers involved). Nativists have postulated that children must be born with extensive knowledge of the structure of human languages. Empiricists, on the other hand, assume that children are able to acquire natural languages using just general-purpose learning mechanisms. They have attempted to show this in neural network models of learning and development, but have by and large failed to convince nativists. Based on the models and results from this thesis and related work, I argue that the persistence of the disagreement can be understood from the fact that empiricists have not provided a satisfactory answer to two fundamental questions:

1. How can it be that children always guess right? That is, from the wide range of logical possibilities, how come children always choose the ones that are consistent with observed constraints on natural languages?
2. How can children learn the class of human languages without specific prior knowledge in the light of all the negative formal learnability results?

Each of these questions leads to a variant of the "argument from the poverty of the stimulus" (see Pullum & Scholz, 2002, for a brief history), which states that much of human language must

be innate because there is insufficient information in the "primary linguistic data". I will briefly review both versions of the argument, and then consider the new twist that evolutionary models as studied in this thesis can give to this old debate.

7.3.1 The Poverty of the Stimulus I

Despite many revisions of theories in the nativist tradition, the argument from the poverty of the stimulus continues to be the prime justification for assuming an extensive, innate Universal Grammar. For example, Jackendoff (2002, p. 69) uses the following quote from Chomsky (1965) to explain the argument:

> *"It seems clear that many children acquire first or second languages quite successfully even though no special care is taken to teach them and no special attention is given to their progress. It also seems apparent that much of the actual speech observed consists of fragments and deviant expressions of a variety of sorts. Thus it seems that a child must have the ability to "invent" a generative grammar that defines well-formedness and assigns interpretations to sentences even though the primary linguistic data that he uses as a basis for this act of theory construction may, from the point of view of the theory he constructs, be deficient in various respects." (Chomsky, 1965, p. 200-1)*

Well-known examples of the miraculous choices that children make concern, among others, *wh*-transformation (i.e. the ways to formulate who, what and where questions), negations, pronouns and quantification. For instance, in example 7.1 (from Crain, 1991), there are complex constraints on whether or not "he" can refer to "the Ninja Turtle".

(7.1) a. The Ninja Turtle danced while he ate the pizza

b. He danced while the Ninja Turtle ate the pizza

c. While he danced the Ninja Turtle ate a pizza

Children know that "he" in sentence (b) cannot refer to the "the Ninja Turtle", and, according to the argument, they need innate syntactic constraints because "there are no data available in the environment corresponding to the kind of negative facts that constraints account for" (Crain, 1991).

Jackendoff (2002, p. 85) uses the following example, from Gruber (1965), to explain what he calls the "Paradox of Language Acquisition":

(7.2) a. Every acorn grew into an oak.

b. Every oak grew out of an acorn.

c. An oak grew out of every acorn.

d. *An acorn grew into every oak.

"Every" in the first three examples quantifies over both the oak and the acorn, while in (d) it quantifies only over the oak, making the sentence uninterpretable. There seems to be no "natural" solution to this idiosyncracy of language. The paradox is that while linguists struggle with the construction, children miraculously always get it right:

> *"The community of linguists, collaborating over many decades, has so far failed to come up with an adequate description of a speaker's knowledge of his or her native language. Yet every normal child manages to acquire this knowledge by the age of ten or so, without reading any linguistics textbooks or going to any conference. How is it that in some sense every single normal child is smarter than the whole community of linguists?" (Jackendoff, 2002, p. 83)*

Chomsky, Crain, Jackendoff and many others have used such examples to argue for the existence of innate, language-specific knowledge. If children always make the right choice, even though clearly sensible logical alternatives exist based on the available evidence, then obviously they must have prior knowledge of the task. Universal Grammar is the theory of that prior knowledge. Pinker & Bloom (1990) and others, have argued that it is "vanishingly unlikely" that this innate knowledge is a side-effect of the evolution of general learning mechanisms; rather, they claim, it must be the result of a gradual process of natural selection. That is, the Universal Grammar is a language-specific adaptation of humans for the use of natural language:

> *"[..] it would be vanishingly unlikely for something that was not designed as a television set to display television programs; the engineering demands are simply too complex. [...] We suggest that human language is a similar case. We are not talking about noses holding up spectacles. Human language is a device capable of communicating exquisitely complex and subtle messages, from convoluted soap opera plots to theories of the origin of the universe. Even if all we knew was that humans possessed such a device, we would expect that it would have to have rather special and unusual properties suited to the task of mapping complex propositional structures onto a serial channel, and an examination of grammar confirms this expectation." (Pinker & Bloom, 1990)*

7.3.2 *The Poverty of the Stimulus - II*

The case for such a view on Universal Grammar is often said to be strengthened by mathematical results from learnability theory. This different version of the argument from the poverty of stimulus starts from the observation that the formalisms that are needed to describe natural language have to be able to represent recursive, hierarchical phrase-structure. Humans can generalise from the sentences they have heard to completely new expressions, many of which might not have been used by anyone ever before. This is known as the "productivity argument" (Chomsky, 1955). In fact, as I discussed in chapter 6, although there is now a wide range of competing grammatical theories, there is also converging evidence that a proper formalism for human language has to be mildly context-sensitive (Joshi *et al.*, 1991). Moreover, there is broad consensus that children have to learn their language from primary linguistic data that is relatively impoverished: children cannot solely rely on negative evidence, semantic information, carefully selected training sentences ("motherese") or statistical cues (e.g. Atkinson, 2001).

It is therefore not unreasonable to study the learnability properties of popular formalisms, and try to derive general results on whether it is possible to learn grammars (of the type that linguists agree are necessary for human language) from primary linguistic data (of the type that psycholinguists believe is available to the child). The first to establish such general results was Mark Gold (1967). He emphasised the interrelatedness of assumptions about the nature of human grammar and the nature of the language acquisition process:

> *"[...] a model of the rules of usage of natural languages must be general enough to include the rules which do occur in existing natural languages. This is a lower bound on the generality of an acceptable linguistic theory. On the other hand, the considerations [on learnability] impose an upper bound on generality: For any language which can be defined within the model there must be a training program, consisting of implicit information, such that it is possible to determine which of the definable language is being presented." (Gold, 1967, p. 448)*

Gold put forward a criterion for learnability and a formal characterisation of the available training data. Gold's learnability criterion, "identification in the limit", is a criterion for a *class of languages*, and not for individual languages per se[2]. He showed, for each of these types of data and for a number of formal classes of languages, whether or not they were identifiable in

[2]With respect to a specific algorithm, it makes sense to ask whether it can learn a specific language. Gold, however, asks whether *any* algorithm can learn all the languages from a given class. Hence, the words "learnable" and "learnability" can mean rather different things when they are used in the context of a specific algorithm, or in the general sense with reference to a class of languages. Unfortunately, much confusion about this distinction exists in the literature (Scholz & Pullum, 2002; Johnson, 2004)

the limit. Most famous is his result that super-finite classes of languages, which include the classes of context-free and of context-sensitive languages, are not learnable from positive data. Similar negative learnability results have been obtained with less restrictive learnability criteria (e.g. Wexler & Culicover, 1980, discussed in Bertolo, 2001). These negative results have led to a consensus among theorists of language acquisition in the generative tradition that human languages are simply not learnable without serious innate constraints:

> *"The basic results of the field [of learnability theory] include the formal, mathematical demonstration that without serious constraints on the nature of human grammar, no possible learning algorithm can in fact learn the class of human grammars." (Wexler, 1999)*

However, after initial pessimism on the learnability of formal grammars, more positive results also emerged. Most notably: (i) Horning (1969, discussed in Bertolo, 2001), which showed that the class of stochastic context-free grammars are identifiable in the limit; (ii) Wexler & Hamburger (1973, discussed in Batali, 2002) which showed that context-free grammars are identifiable in the limit if the input includes the proper semantic information for every sentence; and (iii) Angluin (1980, discussed in Kanazawa, 1998), which showed that non-trivial classes of grammars, that include subsets of context-free and context-sensitive grammars, are identifiable in the limit from positive data.

Encouraged by such positive results, a small number of researchers has worked on designing algorithms that indeed induce grammars from examples (e.g. Wolff, 1982; Stolcke, 1994; Clark, 2001; Klein & Manning, 2002; Adriaans *et al.*, 2002), with some success. How can such positive results be reconciled with the negative learnability results that are quoted so often in the nativist tradition? It is worth considering that question in some detail, since it relates directly to the connectionist and evolutionary models that we will discuss later.

Gold (1967), in fact, is very careful in his discussion of the relevance of his negative results. He discusses three solutions for the learnability problem (i) additional restrictions on the class of possible human grammars, (ii) indirect negative evidence and (iii) a priori restrictions on the class of training samples that a child may expect. We can interpret each of the mentioned positive results as elaborations on these possible solutions. That is, Horning (1969) is variant of solution (iii), because it assumes that the learner can rely on statistical information in the training data; Wexler & Hamburger (1973) is a variant of solution (ii), because it assumes that learner can derive indirect negative evidence from the semantics of training sentences; and Angluin (1980) is variant of solution (i) by identifying non-trivial, but restricted classes.

Much discussion has followed on whether or not the child has negative evidence available, and whether or not parents adapt their infant-directed speech to facilitate language learning. Again a consensus has emerged among nativists that these potential solutions are empirically invalidated and that suggestions of the contrary can be safely ignored. For example, in Bertolo (2001), a recent introduction to "learnability theory" with contributions from 6 different authors, none of the theoretical and algorithmic positive results above, except for Horning's, has even been referred to. Instead, researchers in this field have concentrated on models within a Principles and Parameters framework, in which learning is restricted to setting a small number of parameters. It is important to note, however, that even if Gold's solution (i) is adopted, there is still a significant jump to make to go from restricting the class of possible grammars, to assuming an extensive innate Universal Grammar with "parameter setting" as the only challenge in the acquisition of syntax (Scholz & Pullum, 2002). Gold describes solution (i) as follows:

> *"The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally. Equivalently, we may say that the child starts out with more information than that the language it will be presented is context-sensitive. In particular, the results on learnability from text imply the following: The class of possible natural languages if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality." (Gold, 1967)*

Nothing in these formal results shows that the necessary constraints are language-specific adaptations; they could simply be generic properties of the human brain, or, as in the model of chapter 6, the result of a form of cultural evolution. On *intuitive* grounds, that possibility is dismissed as "vanishingly unlikely". Moreover, also the nature of the available linguistic data is an *empirical* and not a formal issue. Although solid empirical results exist that make many of proposed sources unlikely, the jump in the nativist literature to (always) assume the worst-case scenario of only positive data[3] ("text") is premature[4]. Nevertheless, research that makes different assumptions is often treated with rather unjustified disdain in the learnability literature. For example:

> *"It seems that some psychologists, suspicious of the innateness claims which have provided the intellectual backdrop to so much of the progress in modern linguistics,*

[3] In fact, the assumption is even stronger than just "positive data", because it considers positive data in any order, no matter how bizarre.

[4] It seems even in contradiction with other empirical results; see the discussion on children (not) learning language from television conversations in Pinker (1994, p. 278).

> *have found it difficult to give up on the belief that linguistic environments really do have properties (if only we could identify them) which would enable us to see them as providing a sufficient basis for grammar induction. We can be fairly confident in our conclusions under this heading, but we owe it to the skeptics to provide some justification for this confidence." — (Atkinson, 2001, p. 16)*

Hence, we can conclude that the variant of the argument from the poverty of stimulus that has emerged from the field of learnability theory is based on firm, formal results that show that learning is impossible without the proper and interrelated assumptions on what grammars are possible in the first place, and what primary data is available. In a sense, these results reflect the widely accepted view that "pure empiricism", that is, learning without a bias (starting with a "tabula rasa") is an untenable position (as even anti-nativists emphasise, Elman *et al.*, 1996). From that basis, a much less formal argument can be made that the necessary constraints must be innately specified because it is so unlikely that they derive from general constraints on cognition, communication and learning.

7.3.3 Empiricist Arguments

The challenge to empiricists is to provide an adequate response to this intuitive stance, and to present an alternative theory that is general enough to include observed rules in human languages and restricted enough to allow their acquisition. Empiricists have attempted to show the feasibility of such an alternative account (see e.g. Seidenberg, 1997, for a review). The most prominent approach has been to build fully specified neural network models that can display some realistic linguistic behavior. Rumelhart & McClelland's (1986) English Past Tense model without "rules" and Jeff Elman's (1991) Simple Recurrent Neural Networks to learn (fragments of) context-free languages and their successor models are the most well-known examples.

Elman's models, for instance, are specifically designed to disprove the poverty of the stimulus argument, as is evident from titles like *Learning the unlearnable* (Lewis & Elman, 2001). However, there are some serious theoretical problems with the kind of syntactic patterns these networks can represent in principle (long-distance dependencies, see for instance, Steedman, 2002a), as well as some lack of clarity about what the networks have actually learned[5]. Steedman (2002a), Jackendoff (2002) and many others have laid out unsolved challenges for connectionists for modelling syntactic and semantic phenomena.

[5]Typically, error rates on predicting the next word are reported, but these conflate performance on trivial task, e.g. nouns follow determiners, with performance on syntactically challenging tasks.

More generally, however, the problem with the whole approach of disproving the poverty of the stimulus, is that such models, even when successful, do not do away with language specific innate knowledge, shaped by natural selection in a process of biological evolution (Pinker & Bloom, 1990). They may push back how much language-specific knowledge must be assumed innate, but they still depend on a large number of crucial parameters and architecture design choices: parameters that regulate when to insert a new hidden unit, different activation functions for hidden and output units, mechanisms to update weights and topological relations between hidden units, phonological feature extraction, inflection classes. Presumably, these design choices are all crucial for the observed behaviour.

Therefore, although connectionist models might suggest modifications to the nativist theories, they do not solve the poverty of the stimulus: the primary linguistic data and general purpose learning mechanisms are not sufficient. Hence, even if connectionists would meet all these challenges (and work by Pollack, 1988 and subsequent work, comes a long way; moreover, formal results on Turing equivalence of neural networks, Siegelmann & Sontag, 1991, ensure that it is possible), nativists could simply argue that all they have accomplished is a lower-level implementation of the nativist theory. These architectures and parameters are language specific, and not some random variant of general learning mechanisms; and that infants have the proper innate architectures and parameters for learning language is exactly what nativist theory has been claiming all the time. *They are the Universal Grammar.*

Kirsh (1992), in an early (constructive) critique of connectionist responses to the stimulus poverty arguments, formulates this as follows:

> *"[...] to discover a network that will learn successfully, designers must choose with care the network's architecture, the initial values the weights are set to, the learning rule, and the number of times the data set is to be presented to the network [...]. If such parameters are not controlled for, successful learning is extremely improbable. In thoughtful modelling, these parameters are chosen on the basis of assumptions about the nature of the function the system is to learn. That is, on the basis of assumptions about the task and task domain. Prima facie, then, although the learning mechanism operating on data is a general one, the success of this mechanism depends equally on a set of antecedent choices that seem to be domain specific." (Kirsh, 1992, p. 297)*

7.3.4 Language Evolution

Instead, from the field of language evolution a simple but much more promising answer has been put forward: children are so good at learning language, because languages have adapted to the

idiosyncracies of infants learning strategies:

> *"Human children appear preadapted to guess the rules of syntax correctly, precisely because languages evolve so as to embody in their syntax the most frequently guessed patterns. The brain has co-evolved with respect to language, but languages have done most of the adapting." (Deacon, 1997, p. 122)*

It might be "extremely improbable" that a *random* learning mechanism learns a *random* language successfully, it does not necessarily follow that the learning mechanism therefore has to be adapted to the task (Fodor, 1989; Kirby, 1994; Christiansen, 1994; Deacon, 1997). The probability that a random learning mechanism is successful when learning a language that is shaped by generations that were using the same mechanism, is something entirely different and could be very high. This proposal — in its informal, verbal presentation — has been met with some understandable skepticism by nativists. In response to the quote above, Jackendoff has commented:

> *"But this puts the cart before the horse. Deacon is correct that human languages do not push the envelope of Universal Grammar very much. But our question is: What is this envelope anyway such that languages, however they evolve over time, must conform to it? Given all the differences among the languages of the world, what is it about them that enables children to "guess the rules of syntax" so well? This something, whatever it is, is what is meant by Universal Grammar." (Jackendoff, 2002, p. 81-82)*

This criticism reflects the common idea in generative linguistics that the structure of languages which we observe, directly reflects the structure of the innate "envelope" for language. In other words, this is the idea that the *theory of language universals* is the same as the *theory of language innateness.* If one accepts the Uniformity Assumption, and the restricted notion of UG that comes with it, Jackendoff's stance is correct. For instance, Niyogi & Berwick (1995), Yang (2000) and others have studied models of language change based on parameter-setting learning algorithms. Unsurprisingly, the dynamics in such models are rather straightforward. In a Principles and Parameters framework, the innate "envelope" is so restrictive, that all there is left for the cultural process is to move from one particular setting of the parameters to another, without any qualitative change in the expressivity or learnability of the language (Zuidema, 2003b; see appendix C of this thesis).

However, if the "envelope of Universal Grammar" is less restrictive, can more interesting things happen? That is, if we do not start from the assumption that there is a restrictive, innate UG, but rather allow the cultural transmission process to favor certain types of languages over

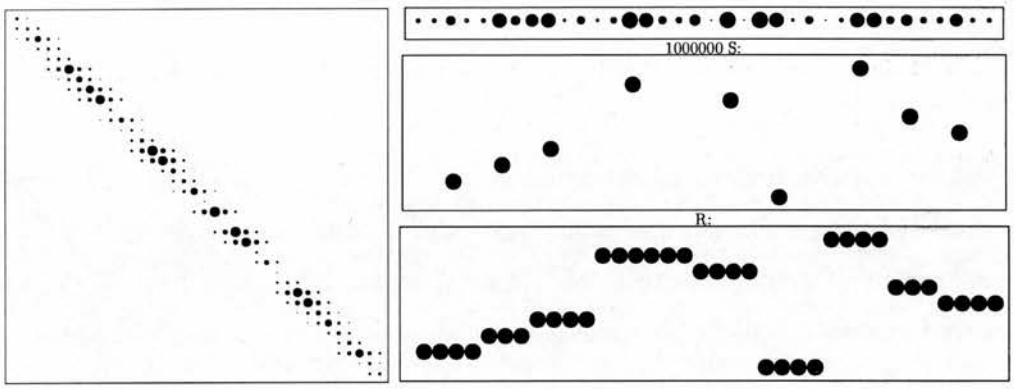
others and actually make qualitative changes in the language, do we observe anything like the process that Deacon describes? This is exactly what the model in chapter 6, and the related Expression–Induction models discussed in that chapter, address. These models show that if we make the obvious assumption that the result of learning in one generation is the output for the next, then the necessary constraints for learnability automatically emerge. Moreover, these models show that the Uniformity Assumption is in some sense a worst case scenario for learnability. If we, in contrast, assume that the innate "envelope" allows a wide range of qualitatively different languages, a process of language adaptation will shape the languages over successive generations to become progressively better learnable.

The fact that languages can change over time and adapt to the language user creates an interesting methodological problem for research on language evolution[6]. Because of this fact, the "appearance of design" cannot, by itself, be taken as evidence for adaptation, although it is used as such in many studies (e.g. Pinker & Bloom, 1990; Jackendoff, 2002; Pinker & Jackendoff, 2004). For instance, Pinker & Jackendoff (2004) present the fact that human vowel perception appears to be different from non-human primate vowel perception as evidence for the view that human hearing has been shaped by natural selection for speech perception. If we accept the idea, however, that languages can adapt to their users, there is a reasonable alternative hypothesis: any arbitrary feature of human hearing, whether or not it has been selected for, will be reflected in the structure of human language because language will adapt to it.

To illustrate this idea I have run the model of chapter 5 with a U matrix where randomly chosen signals are more reliably recognised than others (see figure 7.1a). After running the hill-climbing heuristic, we see in the resulting language, figure 7.1b, that language reflects these arbitrary features of the agents' perception. Language adapts to the language user, rather than the other way around. If we interpret the hill climbing in the model of chapter 5 as learning, there is a perfectly valid explanation for the match between language and user that involves no biological evolution (see Kirby, 1999, for an analogous argument about syntactic patterns in language and cognitive constraints on parsing).

We can conclude that in addition to (i) a theory on what must be innate given the available primary linguistic data, we need (ii) a sophisticated theory on how the structure of the languages

[6]It also poses a methodological problem for language acquisition research: "Children must discover the rules that generate an infinite set, with only a finite sample. They evidently possess additional language-learning abilities that enable them to organize their language without explicit guidance. These abilities diminish with age and may be biologically based. However, scientific efforts to isolate them experimentally encounter a methodological complication: given that today's languages were acquired by children in the past, language input to children already includes products of innate biases. It is therefore difficult to determine whether any particular linguistic element observed in a child's language is inborn or derived." (Saffran *et al.*, 2001; references omitted).



(a) The heterogeneous **U** matrix, with randomly chosen diagonal values.

(b) From top to bottom: **U** matrix diagonal values, **S** matrix and $\mathbf{R}^T$ matrix. **S** and $\mathbf{R}^T$ at $t = 1000000$ (at equilibrium).

Figure 7.1: Results from a simulation with the model of chapter 5, a heterogeneous **U** matrix in the "local optimisation of a deterministic lexicon" condition. The resulting language reflects arbitrary features of the agents' perception.

of the world has emerged given the (innate) learning abilities, the cognitive, articulatory and perceptual features of humans and the processes of cultural transmission (Kirby, 1999). It is clear that the model discussed in chapter 6 is a poor version of such a theory. The model was designed to be as simple as possible and to study the phenomenon of language adaptation *in abstracto* and its relevance for arguments from learnability theory.

A likely conclusion from this new line of research is that with "general" learning algorithms that did not evolve for language, only languages that are structurally quite *different* from contemporary, human languages will emerge. The door is therefore open for language-specific adaptations as imagined in the nativist tradition. However, some of their critics have emphasised that they have no principled objections against such adaptations (Elman *et al.*, 1996). For instance:

> *"Once it finally appeared on the planet, it is quite likely that language itself began to apply adaptive pressure on the organization of the human brain [...]." (Bates & Goodman, 1999)*

The picture that emerges is one where children's learning mechanisms shape the languages over a number of generations, and natural selection shapes the learning mechanisms. Senghas *et al.* (2004) recently phrased a very similar conclusion, based on many years of empirical research on the emergence of a new sign language in a school for the deaf in Nicaragua:

> *"In this way, evolutionary pressures would shape children's language learning (and now, language-building) mechanisms to be analytical and combinatorial. On the*

> *other hand, once humans were equipped with analytical, combinatorial learning mechanisms, any subsequently learned languages would be shaped into discrete and hierarchically organized systems." (Senghas et al., 2004, p.1782)*

Such a perspective turns the argument from the poverty of the stimulus on its head. Language – by virtue of it being a culturally transmitted code – is necessarily learnable. Some of its features, such as hierarchical phrase-structure, do not present a *problem* for learning, but rather a *solution* for cultural transmission through a bottleneck. Children only appear to have prior knowledge, because they happen to make the same arbitrary choices as all the generations before them; they appear to have been adapted for language, because the sounds, meanings and rules of language have been shaped by the learning and usage of previous generations. However, the learning mechanisms might in turn have been shaped by natural selection such that the complex outcome of the cultural evolution is biological adaptive. Perhaps it is not totally inconceivable that such a view will help to bring the ongoing empiricism–nativism debate to an end.

7.4 Implications for Biology

The most obvious implication for biology of the work described in this thesis, is for issues in the evolution of animal communication. The combinatorial phonology in the songs of birds, cetaceans and gibbons (Ujhelyi, 1996), as well as the topology preservation in vervet alarm calls (Seyfarth & Cheney, 1997) and rudimentary compositional semantics in Campbell monkey calls (Zuberbühler, 2002), bee dances (von Frisch, 1965, 1974) and perhaps gibbon calls (Ujhelyi, 1996), might all have evolved the same way as they did in human language. Similarly, the syntactic patterns and recursion observed in the songs of some birds, might have the same origin as such patterns in human language. The models of chapters 4, 5 and 6, or adaptations of these models, might therefore be applicable to evolutionary questions about communication in all these other species.

The observation that languages themselves can evolve culturally points at another possible connection with evolutionary biology. A whole tradition exists of drawing parallels between biological evolution and language change, going back at least to Charles Darwin. Darwin was influenced by observations from historical linguistics before he formulated his theory of evolution by natural selection (Darwin, 1859), and wrote about the parallels in his later work. For instance, in the *Descent of Man*, Darwin writes (as quoted in Mesoudi *et al.*, 2004):

> *"A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the*

upper hand, and they owe their success to their own inherent virtue." (Darwin, 1871, p.91)

In modern terminology, we can draw parallels between species and language, between the gene and the words and rules[7] of a language, between speciation and language birth, and between extinction and language death and so-forth. This is not just an amusing curiosity; potentially, the tools and concepts of evolutionary biology can be used to analyse data from historical linguistics and linguistic typology. For linguists, one benefit is that the field of mathematical modelling of evolution (as discussed in chapter 2) is much more advanced than mathematical modelling of language change (e.g. Niyogi & Berwick, 1995; Yang, 2000). For biologists, there could be benefits from applying and extending their tools to an alternative domain, where much empirical data is available (and more and more is easily accessible through the internet).

Finally, a third possible connection between the models in this thesis and issues in biology, concerns the origins of the genetic code. In descriptions of genetics, biologists have always used linguistic terminology, such as code, information, expression, translation, transcription, "language of the genes", and so-forth. But the analogy goes deeper than that. Through a by now well understood code, almost universal for life on earth, specific triplets of DNA nucleotides (the building blocks of genes) code for specific amino-acids (the building blocks of proteins). The origins of the genetic code are, like the origins of human language, still largely an open issue and there are many parallels to be drawn (Maynard Smith & Szathmáry, 1995; Nick Barton, p.c.). The DNA codon, or the transfer-RNA "copy" of it, is the analogue of a word, and the amino-acids of its meaning.

Szathmáry (1993) has proposed a scenario for its origin, which involves an earlier stage where amino-acids help *ribozymes* (RNA based enzymes) to catalyse reactions, and where RNA "handles" are the precursors of modern transfer-RNA. These handles attach to specific amino-acids, and help to position the amino-acid precisely on the ribozyme. Szathmáry imagines that at this stage each amino-acid (the "meaning") can get attached to multiple RNA handles ("words"), as in models (as for instance, in chapter 5) of the cultural evolution of language where each meaning can be expressed by many different words (signals). In this scenario, different amino-acid/RNA-handle combinations are in competition with each other. Eventually, a specific handle gets established for each amino-acid, and a precursor genetic code emerges. Interestingly, there is even a form of "topology preservation", where similar codons tend to code for similar amino-acids.

[7]The best analogy with the gene is probably the lexical entry (word) and its associated meaning and syntactic category or "supertag" in lexicalised frameworks (e.g. Gamut, 1991; Steedman, 2000; Joshi, 2004).

7.5 Future Work

This thesis provided an exploration of a complete and formal scenario of the evolution of language. Much more work can and should be put into analysing and extending the models proposed in chapters 4, 5 and 6. An important extension to chapter 4 would be to provide a good measure of the degree to which a signal repertoire is combinatorial. In chapters 4 and 5, I made the assumption that with a combinatorial or compositional E-language, these features can more easily evolve in the I-language. This assumption should be tested in follow-up models.

The model in chapter 6 should be replaced by a model with a much more robust learning algorithm, that can be parametrised such that the evolutionary dynamics of a wide set of possible learning strategies can be explored. I believe such a new model could be both important for theories of language evolution, and much strengthen the discussed implications for linguistics and biology. Unfortunately, the unsupervised learning of grammar is a major problem forms a whole project in itself. I have started to work on a new algorithm for unsupervised grammar learning, that unlike the algorithm in chapter 6 works with a stochastic grammar formalism (stochastic tree substitution grammars, which should make the learning more robust), includes semantic representation (based on the lambda calculus, which should make the definition of payoff more straightforward) and integrates, like Batali (2002), learning with parsing (based on memoised, left-corner parsing techniques).

Is a serious scientific investigation of language origins feasible? As I discussed at the start of this thesis, many scholars worry that the problem is underdetermined, that is, that there will always be many explanations consistent with the scarce empirical facts. I hope to have shown in this thesis that evolutionary biology and linguistics bring enough formal constraints on evolutionary explanations to evaluate (and reject) many current proposals, and to define clear challenges for mathematical and computational modellers. Whether the problem really is principally underdetermined is an open issue, but I feel there is every reason to try to meet these challenges and work out rigorous scenarios in the framework sketched.

Ultimately, a detailed understanding of how language evolved will depend on a detailed understanding of how language works: How can the biological hardware of the brain process language? How does a child acquire the knowledge of her native language? How does the structure of natural languages depend on these learning and processing mechanisms? These are formidable challenges. If evolutionary linguistics can contribute anything to answering these questions, it will prove to be not only a fascinating but also a worthwhile exercise.

References

ABBOTT, B. (1999). The formal approach to meaning. *Journal of Foreign Languages (Shanghai)* **119**, 2–20.

ADRIAANS, P., FERNAU, H. & VAN ZAANEN, M., eds. (2002). *Grammatical Inference: Algorithms and Applications (Proceedings of the 6th International Colloquium on Grammatical Inference*, vol. 2484 of *Lecture Notes in Computer Science*, Berlin. Springer.

ANGLUIN, D. (1980). Inductive inference of formal languages from positive data. *Information and Control* **21**, 46–62.

ARCADI, A. (1996). Phrase structure of wild chimpanzee pant hoots: patterns of production and interpopulation variability. *Am. J. Primatol* **39**, 159–178.

ATKINSON, M. (2001). Learnability and the acquisition of syntax. In: Bertolo (2001), chap. 2.

BARTON, G. E. & BERWICK, R. C. (1987). *Computational Complexity and Natural Language*. Cambridge, MA: MIT Press.

BARTON, N. & PARTRIDGE, L. (2000). Limits to natural selection. *BioEssays* **22**, 1075–1084.

BARTON, N. & TURELLI, M. (1991). Natural and sexual selection on many loci. *Genetics* **127**, 229–255.

BARTON, N. & ZUIDEMA, W. (2003). Evolution: the erratic path towards complexity. *Current Biology* **13**, 649–651.

BATALI, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In: *Artifical Life IV* (Brooks, R. & Maes, P., eds.), pp. 160–171. Cambridge, MA: MIT Press.

BATALI, J. (1998). Computational simulations of the emergence of grammar. In: Hurford *et al.* (1998).

BATALI, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe (2002b).

BATES, E. & GOODMAN, J. C. (1999). On the emergence of grammar from the lexicon. In: *The emergence of language* (MacWhinney, B., ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

VON BEKESY, G. (1960). *Experiments in hearing*. New York, NY: McGraw-Hill.

BELPAEME, T. (2001). *Factors influencing the origins of colour categories*. Ph.D. thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel.

BERTOLO, S., ed. (2001). *Language Acquisition and Learnability*. Cambridge University Press.

BICKERTON, D. (1990). *Language and Species*. Chicago, IL: University of Chicago Press.

BICKERTON, D. (2003a). Language evolution without evolution. *Behavioral and Brain Sciences* **26**, 669–670.

BICKERTON, D. (2003b). Symbol and structure: a comprehensive framework for language evolution. In: Christiansen & Kirby (2003a), pp. 77–93.

BOD, R. (1998). *Beyond Grammar: An experience-based theory of language*. Stanford, CA: CSLI.

BOD, R. (2003). An efficient implementation of a new DOP model. In: *Proceedings EACL'03*.

DE BOER, B. (1999). *Self Organisation in Vowel Systems*. Ph.D. thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel.

DE BOER, B. (2000). Self organization in vowel systems. *Journal of Phonetics* **28**, 441–465.

DE BOER, B. (2001). *The origins of vowel systems*. Oxford, UK: Oxford University Press.

DE BOER, B. & ZUIDEMA, W. (2003). Phonemic coding: Optimal communication under noise? In: *Proceedings of the Workshop on Language Evolution and Computation* (Kirby, S., ed.). 15th European Summer School in Logic Language and Information (ESSLLI).

BOERLIJST, M. & HOGEWEG, P. (1991). Self-structuring and selection: Spiral waves as a substrate for prebiotic evolution. In: *Artificial Life II* (Langton, C., Tayler, C., Farmer, J. & Rasmussen, S., eds.), pp. 255–276.

BOGERT, B. P., HEALY, M. J. & TUKEY, J. W. (1963). The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and shape cracking. In: *Time Series Analysis* (Rosenblatt, M., ed.), pp. 209–243. New York, NY: J. Wiley.

BOTHA, R. (2003). *Unravelling the Evolution of Language*. Oxford: Elsevier.

BOYD, R. & RICHERSON, P. (1985). *Culture and the Evolutionary Process*. Chicago, IL: Chicago University Press.

BRIGHTON, H. (2002). Compositional syntax from cultural transmission. *Artificial Life* **8**.

BRIGHTON, H. (2003). *Simplicity as a Driving Force in Linguistic Evolution*. Ph.D. thesis, Theoretical and Applied Linguistics, University of Edinburgh.

BRISCOE, T. (2000a). Evolutionary perspectives on diachronic syntax. In: *Diachronic Syntax: Models and Mechanisms* (Pintzuk, S., Tsoulas, G. & Warner, A., eds.). Oxford, UK: Oxford University Press.

BRISCOE, T. (2000b). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language* **76**.

BRISCOE, T. (2002a). Grammatical acquisition and linguistic selection. In: Briscoe (2002b).

BRISCOE, T., ed. (2002b). *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

BRISCOE, T. (2003). Grammatical assimilation. In: Christiansen & Kirby (2003a), pp. 317–337.

BUSZKOWSKI, W. & PENN, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica* **49**, 431–454.

BYBEE, J. L. (2003). Mechanisms of change in grammaticization: the role of frequency. In: *Handbook of Historical Linguistics* (Janda, R. & Joseph, B., eds.). Oxford: Blackwell. to appear.

CANGELOSI, A. & PARISI, D., eds. (2002). *Simulating the Evolution of Language*. London: Springer Verlag.

CARLSON, R., GRANSTRÖM, B. & FANT, G. (1970). Some studies concerning perception of isolated vowels. In: *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 2-3, pp. 19–35. Stockholm, Sweden: Royal Institute of Technology.

CAVALLI-SFORZA, L. & FELDMAN, M. (1983). Paradox of the evolution of communication and of social interactivity. *Proc. Nat. Acad. Sci. USA* **80**, 2017–2021.

CAVALLI-SFORZA, L. L. & FELDMAN, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

CHIBA, T. & KAJIYAMA, M. (1958). *The Vowel: Its Nature and Structure*. Tokyo: Phonetic Society of Japan.

CHIERCHIA, G. & MCCONNELL-GINET, S. (1990). *Meaning and Grammar*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1955). *Logical Structure of Linguistic Theory*. Plenum.

CHOMSKY, N. (1957). *Syntactic Structures*. The Hague: Mouton.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1972). *Language and mind*. Harcourt, Brace and World. Extended edition.

CHOMSKY, N. (1975). *Reflections on Language*. New York: Pantheon.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

CHOMSKY, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

CHOMSKY, N. (2002). Paper presented at the fourth international conference on the evolution of language. Harvard University.

CHOMSKY, N. & HALLE, M. (1968). *The sound pattern of English*. New York, NY: Harper & Row.

CHRISTIANSEN, M. H. (1994). *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. Ph.D. thesis, University of Edinburgh, Scotland.

CHRISTIANSEN, M. H. & KIRBY, S., eds. (2003a). *Language Evolution*. Oxford, UK: Oxford University Press.

CHRISTIANSEN, M. H. & KIRBY, S. (2003b). Language evolution: Consensus and controversies. *Trends in Cognitive Science* **7**, 300–307.

CHRISTIANSEN, M. H. & KIRBY, S. (2003c). Language evolution: The hardest problem in science? In: Christiansen & Kirby (2003a), pp. 1–15.

CLAHSEN, H. (1999). Lexical entries and rules of language: A multidisciplinary study of german inflection. *Behavioral and Brain Sciences* **22**.

CLARK, A. (2001). *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, University of Sussex.

COMRIE, B. (1981). *Language Universals and Linguistic Typology*. Basil Blackwell.

COOK, V. (1993). *Linguistics and second language acquisition*. Macmillan.

COREN, S., WARD, L. M. & ENNS, J. T. (1979/1994). *Sensation and Perception*, vol. Fort Worth, TX. Harcourt Brace.

COYNE, J., BARTON, N. & TURELLI, M. (2000). Is Wright's shifting balance process important in evolution? *Evolution* **54**, 306–317.

CRAIN, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences* **14**, 597–611.

CROW, J. F. (1999). Hardy, Weinberg and language impediments. *Genetics* **152**, 821–825.

DARWIN, C. (1859). *The Origin of Species – by means of natural selection or the preservation of favoured races in the struggle for life*. London: Murray. (this edition, New York: The New American Library, 1958).

DARWIN, C. (1871). *The Descent of Man, and selection in relation to sex*. London: John Murray. Reprinted in 1981 by Princeton University Press.

DAWKINS, R. (1976). *The Selfish Gene*. Oxford University Press. This edition 1989.

DAWKINS, R. (1982). *The Extended Phenotype*. Oxford: Oxford University Press.

DAWKINS, R. & KREBS, J. R. (1978). Animal signals: information or manipulation? In: *Behavioural ecology: an evolutionary approach* (Krebs, J. R. & Davies, N. B., eds.). Oxford, UK: Blackwell Scientific Publications.

DE BEULE, J., VAN LOOVEREN, J. & ZUIDEMA, W. (2002). Grounding formal syntax in an almost real world. Tech. rep., Vrije Universiteit Brussel, AI Memo 02-03.

DE JONG, E. D. (2000). *Autonomous Formation of Concepts and Communication*. Ph.D. thesis, Vrije Universiteit Brussel AI-lab.

DEACON, T. (1997). *Symbolic species, the co-evolution of language and the human brain*. The Penguin Press.

DEACON, T. (2000). Evolutionary perspectives on language and brain plasticity. *Journal of Communication Disorders* **33**, 273–290.

DESSALLES, J.-L. (1998). Altruism, status, and the origin of relevance. In: Hurford *et al.* (1998).

DEUCHAR, M. (1996). Spoken language and sign language. In: *Handbook of Human Symbolic Evolution* (Lock, A. & Peters, C. R., eds.). Oxford, UK: Clarendon Press.

DOBZHANSKY, T. (1937). *Genetics and the Origin of Species*. Columbia University Press.

DONALD, M. (1991). *Origins of the Modern Mind*. Cambridge, MA: Harvard University Press.

DOUPE, A. J. & KUHL, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience* **22**, 567–631.

DUNBAR, R. (1998). Theory of mind and the evolution of language. In: Hurford *et al.* (1998).

DUNBAR, R. (2003). The origin and sunsequent evolution of language. In: Christiansen & Kirby (2003a), pp. 219–234.

EIGEN, M. (1971). Self-organization of matter and the evolution of biological macro-molecules. *Naturwissenschaften* **58**, 465–523.

ELMAN, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* **7**, 195–225.

ELMAN, J. L., BATES, E. A., JOHNSON, M. H., KARMILOFF-SMITH, A., PARISI, D. & PLUNKETT, K. (1996). *Rethinking Innateness. A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

FANT, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton & Co.

FISHER, R. A. (1922). On the dominance ratio. *Proc Roy Soc Edin* **42**, 321–431.

FISHER, R. A. (1930). *The Genetical Theory of Natural Selection.*. Oxford, UK: Clarendon Press.

FITCH, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Science* **4**, 258–267.

FITCH, W. T. (2004). Kin selection and "mother tongues": A neglected component in language evolution. In: *Evolution of Communication Systems: A Comparative Approach* (Oller, K. & Griebel, U., eds.), pp. 275–296. Cambridge, MA: MIT Press.

FITCH, W. T. & HAUSER, M. D. (2002). Unpacking "honesty": Vertebrate vocal production and the evolution of acoustic signals. In: *Acoustic Communication* (Simmons, A., Fay, R. & Popper, A., eds.), vol. 16, pp. 65–137. New York, NY: Springer.

FITCH, W. T. & HAUSER, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science* **303**, 377–380.

FODOR, J. D. (1989). Learning the periphery. In: *Learnability and linguistic theory* (Matthews, R. J. & Demopoulos, W., eds.), pp. 129–154. Dordrecht, the Netherlands: Kluwer.

FRANK, S. A. (1998). *Foundations of Social Evolution*. Princeton University Press.

FREGE, G. (1923). *Logische Untersuchungen (Dritter Teil: Gedankenfüge, Beiträge zur Philosophie des Deutschen Idealismus vol. III, pp. 36-51)*. Oxford, UK: Basil Blackwell. Translated by P.T. Geach and R.H. Stoothoff as *Logical Investigations, Part III: Compound Thoughts*, 1977, pp. 55-78.

FRIEDERICI, A. D. (2004). Processing local transitions versus long-distance syntactic hierarchies. *Trends in Cognitive Sciences* **8**, 245–247.

VON FRISCH, K. (1965). *Tanzsprache und Orientierung der Bienen*. Berlin: Springer-Verlag.

VON FRISCH, K. (1974). Decoding the language of the bee. *Science* **185**, 663–668.

GAMUT, L. (1991). *Logic, language and meaning*, vol. 2. The University of Chicago Press.

GARDNER, A. & WEST, S. (2004). Spite and the scale of competition. *Journal of Evolutionary Biology* in press.

GARDNER, R. & GARDNER, B. (1969). Teaching sign language to a chimpanzee. *Science* **165**, 664–672.

GAZDAR, G. (1981). Unbounded dependencies and coordinate structure. *Linguistic Inquiry* **12**, 155–184. Reprinted in Walter J. Savitch, Emmon Bach, William Marsh and Gila Safran-Naveh, eds. (1987), *The Formal Complexity of Natural Language* Dordrecht: Reidel, pp. 183-226.

GAZDAR, G. & PULLUM, G. (1981). Subcategorization, constituent order and the notion of "head". In: *The Scope of Lexical Rules* (Moortgat, M., van der Hulst, H. & Hoekstra, T., eds.), pp. 107–123. Dordrecht, Holland: Foris.

GOLD, E. M. (1967). Language identification in the limit. *Information and Control (now Information and Computation)* **10**, 447–474.

GOLDBERG, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. (CTLC) Cognitive Theory of Language and Culture Series. Chicago, IL: The University of Chicago Press.

GOLDENTHAL, W. D. (1994). *Statistical Trajectory Models for Phonetic Recognition*. Ph.D. thesis, MIT, Department of Aeronautics and Astronautics.

GRAFEN, A. (1979). The hawk-dove game played between relatives. *Animal Behaviour* **27**, 905–907.

GRAFEN, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology* **144**, 517–546.

GRAFEN, A. (2003). Fisher the evolutionary biologist. *Journal of the Royal Statistical Society: Series D (The Statistican)* **52**, 319–329.

GRUBER, J. (1965). *Studies in Lexical Relations*. Ph.D. thesis, MIT. Repr. in Gruber, Lexical Structures in Syntax and Semantics, Amsterdam.

GUENTHER, F. H. & GJAJA, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustic Society of America* **100**, 1111–1121.

HALDANE, J. B. S. (1932). *The causes of evolution*. New York: Longmans.

HAMILTON, W. (1964a). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology* **7**, 1–16.

HAMILTON, W. (1964b). The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology* **7**, 17–52.

HAMILTON, W. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature* **228**, 1218–20.

HARDY, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28**, 49–50.

HARLEY, C. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology* **89**, 611–33.

HARNAD, S., ed. (1987). *Categorical Perception: the groundwork of cognition*. Cambridge, UK: Cambridge University Press.

HASHIMOTO, T. & IKEGAMI, T. (1996). The emergence of a net-grammar in communicating agents. *BioSystems* **38**, 1–14.

HAUSER, M. D. (1996). *The Evolution of Communication*. Cambridge, MA: Bradford/MIT Press.

HAUSER, M. D. (2001). What's so special about speech? In: *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler* (Dupoux, E., ed.). Cambridge, MA: MIT Press.

HAUSER, M. D., CHOMSKY, N. & FITCH, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579.

HAUSER, M. D. & FITCH, W. T. (2003). What are the uniquely human components of the language faculty? In: Christiansen & Kirby (2003a), pp. 317–337.

HEYES, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences* **21**, 101–134.

HIGGINBOTHAM, J. (1997). GB Theory: An introduction. In: *the Handbook of Logic and Language* (van Benthem, J. F. A. K. & ter Meulen, G. B. A., eds.). Amsterdam: Elsevier.

HILL, W. & ROBERTSON, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**.

HINTON, G. E. & NOWLAN, S. J. (1987). How learning can guide evolution. *Complex systems* **1**, 495–502.

HINTON, L., NICHOLS, J. & OHALA, J. J., eds. (1995). *Sound Symbolism*. Cambridge, UK: Cambridge University Press.

HOCKETT, C. (1960). The origin of speech. *Scientific American* **203**, 88–111.

HORNING, J. (1969). *A study of grammatical inference*. Ph.D. thesis, Computer Science Dep., Stanford University.

VON HUMBOLDT, W. (1836). *On Language*. Texts in German Philosophy. Cambridge, UK: Cambridge University Press. Translated from the German by Peter Heath. This edition 1988.

HURFORD, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222.

HURFORD, J. R. (2000). Social transmission favours linguistic generalization. In: Knight *et al.* (2000).

HURFORD, J. R. (2002a). Expression / Induction models of language. In: Briscoe (2002b).

HURFORD, J. R. (2002b). The roles of expression and representation in language evolution. In: Wray (2002).

HURFORD, J. R. (2003). The neural basis of predicate-argument structure. *Behavioral and Brain Sciences* **26**, 261–283.

HURFORD, J. R., STUDDERT-KENNEDY, M. & KNIGHT, C., eds. (1998). *Approaches to the evolution of language: social and cognitive bases*. Cambridge, UK: Cambridge University Press.

HUYBRECHTS, R. (1984). The weak inadequacy of context-free phrase structure grammars. In: *Van Periferie naar Kern* (de Haan, G., Trommelen, M. & Zonneveld, W., eds.). Foris.

JACKENDOFF, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Science* **3**.

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

JÄGER, G. (2005). Evolutionary game theory for linguists. a primer. Tech. rep., Stanford University and University of Potsdam.

JANSSEN, T. M. V. (1997). Compositionality (with an appendix by Barbara H. Partee). In: *the Handbook of Logic and Language* (van Benthem, J. F. A. K. & ter Meulen, G. B. A., eds.). Amsterdam: Elsevier.

JOHNSON, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science* **71**, 571–592.

JOSHI, A. & SAKAR, A. (2003). Tree adjoining grammars and their application to statistical parsing. In: *Data-Oriented Parsing* (Bod, R., Scha, R. & Sima'an, K., eds.), pp. 253–282. Chicago, IL: CSLI Publications, University of Chicago Press.

JOSHI, A., VIJAY-SHANKER, K. & WEIR, D. (1991). The convergence of mildly context-sensitive grammar formalisms. In: *Foundational issues in natural language processing* (Sells, P., Shieber, S. & Wasow, T., eds.), pp. 21–82. Cambridge MA: MIT Press.

JOSHI, A. K. (1985). How much context-sensitivity is required to provide reasonable structural descriptions: Tree-adjoining grammars. In: *Natural Language Parsing: Psycholinguistic, Computational and Theoretical Perspectives* (Dowty, D., Karttunen, L. & Zwicky, A., eds.), pp. 206–350. New York: Cambridge University Press.

JOSHI, A. K. (2004). Starting with complex primitives pays off: complicate locally, simplify globally. *Cognitive Science* **28**, 637–668.

KANAZAWA, M. (1998). *Learnable Classes of Categorial Grammars*. Stanford CA: CSLI Publications.

KAPLAN, F. (2000). *L'émergence d'un lexique dans une population d'agents autonome*. Ph.D. thesis, Université Paris 6, Sony CSL-Paris.

KAPLAN, R. & BRESNAN, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In: *The Mental Representation of Grammatical Relations* (Bresnan, J., ed.), chap. 4, pp. 173–281. Cambridge, MA: MIT Press.

KATZ, J. J. & POSTAL, P. M. (1964). *An Integrated Theory of Linguistic Descriptions*. Cambridge, MA: MIT Press.

KAY, P. & FILLMORE, C. (1999). Grammatical constructions and linguistic generalizations. *Language* **75**, 1–33.

KIRBY, S. (1994). Adaptive explanations for language universals: A model of hawkins' performance theory. *Sprachtypologie und Universalienforshung* **47**, 186–210.

KIRBY, S. (1998). Fitness and the selective adaptation of language. In: Hurford *et al.* (1998).

KIRBY, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford University Press.

KIRBY, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: Knight *et al.* (2000).

KIRBY, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* **5**, 102–110.

KIRBY, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe (2002b).

KIRBY, S. (2002b). Natural language from artificial life. *Artificial Life* **8**, 185–215.

KIRBY, S. & HURFORD, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In: *Proceedings 4th European Conference on Artificial Life* (Husbands, P. & Harvey, I., eds.), pp. 493–502. Cambridge, MA: MIT Press.

KIRBY, S. & HURFORD, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In: Cangelosi & Parisi (2002), chap. 6, pp. 121–148.

KIRKPATRICK, M., JOHNSON, T. & BARTON, N. (2002). General models of multilocus evolution. *Genetics* **161**, 1727–50.

KIRSH, D. (1992). PDP learnability and innate knowledge of language. In: *Connectionism: theory and practice* (Davis, S., ed.), pp. 297–322. Oxford University Press.

KLEIN, D. & MANNING, C. D. (2002). A generative constituent-context model for improved grammar induction. In: *Proceedings of the 40th Annual Meeting of the ACL*.

KLEIN, W. & PERDUE, C. (1997). The basic variety, or: Couldn't language be much simpler? *Second Language Research* **13**, 301–347.

KNIGHT, C., HURFORD, J. R. & STUDDERT-KENNEDY, M., eds. (2000). *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*. Cambridge University Press.

KOHONEN, T. (1988). The "neural" phonetic typewriter. *Computer* **21**, 11–22.

KOMAROVA, N., NIYOGI, P. & NOWAK, M. (2001). The evolutionary dynamics of grammar acquisition. *J. Theor. Biology* **209**, 43–59.

KOMAROVA, N. L. & NIYOGI, P. (2004). Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artificial Intelligence* **154**, 1–42.

KOMAROVA, N. L. & NOWAK, M. A. (2001). The evolutionary dynamics of the lexical matrix. *Bull. Math. Biol* **63**, 451–485.

KOMAROVA, N. L. & NOWAK, M. A. (2003). Language, learning and evolution. In: Christiansen & Kirby (2003a), pp. 317–337.

KUHL, P., WILLIAMS, K., LACERDA, F., STEVENS, K. & LINDBLOM, B. (1992). Linguistic experience alters phonetic perception in infants by 6 month of age. *Science* **255**, 606–608.

LAI, C., FISHER, S., HURST, J., VARGHA-KHADEM, F. & MONACO, A. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–23.

LANGACKER, R. (1987). *Foundations of Cognitive Grammar*. Stanford: Standford University Press.

LEVELT, W. & WHEELDON, L. (1994). Do speakers have access to a mental syllabary? *Cognition* **50**, 239–69.

LEWIS, D. K. (1969). *Convention: a Philosophical Study*. Cambridge, MA: Harvard University Press.

LEWIS, D. K. (1972). General semantics. In: *Semantics of Natural Language* (Davidson, D. & Harman, G., eds.), pp. 169–218. Dordrecht, Holland: Reidel. Reprinted in Partee, Barbara, ed. 1976. Montague Grammar. New York: Academic Press.

LEWIS, J. & ELMAN, J. (2001). A connectionist investigation of linguistic arguments from the poverty of the stimulus: learning the unlearnable. In: *Proceedings of the 23d Annual Conference of the Cognitive Science Society* (Moore, J. & Stenning, K., eds.), pp. 552–557. Mahway, NJ: Lawrence Erlbaum.

LEWONTIN, R. C. (1990). How much did the brain have to change for speech? *Behavioral and Brain Sciences* **13**, 740–741. Peer commentary on Pinker & Bloom, 1990.

LEWONTIN, R. C. (1998). The evolution of cognition: Questions we will never answer. In: *An invitation to cognitive science* (Scarborough, D. & Sternberg, S., eds.), vol. 4: Methods, models, and conceptual issues. Cambridge, MA: MIT Press.

LIBERMAN, A., COOPER, F., SHANKWEILER, D. & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review* **74**, 431–461.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. & GRIFFITH, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**, 358–368.

LIEBERMAN, P. (1984). *The biology and evolution of language*. Cambridge, MA: Harvard University Press.

LIEBERMAN, P. (2003). Motor control, speech and the evolution of human language. In: Christiansen & Kirby (2003a), pp. 255–271.

LIEVEN, E., BEHRENS, H., SPEARES, J. & TOMASELLO, M. (2003). Early syntactic creativity: a usage-based approach. *J. Child Lang.* **30**, 333–370.

LILJENCRANTS, J. & LINDBLOM, B. (1972). Numerical simulations of vowel quality systems: the role of perceptual contrast. *Language* **48**, 839–862.

LINDBLOM, B., MACNEILAGE, P. & STUDDERT-KENNEDY, M. (1984). Self-organizing processes and the explanation of phonological universals. In: *Explanations for Language Universals* (Butterworth, B., Comrie, B. & Dahl, O., eds.), pp. 181–203. Berlin: Mouton.

MACNEILAGE, P. F. & DAVIS, B. L. (2000). On the origin of internal structure of word forms. *Science* **288**, 527–531.

MANNING, C. & SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

MANTAKAS, M., SCHWARTZ, J. & ESCUDIER, P. (1986). Modèle de prédiction du deuxième formant effectif F2'– application à l'étude de la labialité des voyelles avant du français. In: *Proceedings of the 15th journées d'étude sur la parole*, pp. 157–161. Société Française d'Acoustique.

MASATAKA, N. (1987). The perception of sex-specificity in the long calls of the tamarin (saguinnes labiatus labiatus). *Ethology* **76**, 56–64.

MAYNARD SMITH, J. (1964). Group selection and kin selection. *Nature* **201**, 1145–1147.

MAYNARD SMITH, J. (1965). The evolution of alarm calls. *The American Naturalist* **99**, 59–63.

MAYNARD SMITH, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, England.

MAYNARD SMITH, J. & HAIGH, J. (1974). The hitch-hiking effect of a favourable gene. *Genet.Res.* **23**, 23–35.

MAYNARD SMITH, J. & PRICE, G. R. (1973). The logic of animal conflict. *Nature* **246**, 15–18.

MAYNARD SMITH, J. & SZATHMÁRY, E. (1995). *The major transitions in evolution*. Oxford: W.H. Freeman.

MESOUDI, A., WHITEN, A. & LALAND, K. (2004). Perspective: is human cultural evolution darwinian? evidence reviewed from the perspective of the origin of species. *Evolution* **58**, 1–11.

MITANI, J. C. & MARLER, P. (1989). A phonological analysis of male gibbon singing behavior. *Behaviour* **109**, 20–45.

MITCHENER, G. & NOWAK, M. A. (2002). Competitive exclusion and coexistence of universal grammars. *Bull Math Biol* **65**, 67–93.

NASH, J. F. (1950). Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **36**, 48–49.

NEWMEYER, F. J. (2003). What can the field of linguistics tell us about the orgins of language? In: Christiansen & Kirby (2003a), pp. 58–76.

NIYOGI, P. (1998). *The informational complexity of learning*. Boston, MA: Kluwer.

NIYOGI, P. (2002). Theories of cultural evolution and their applications to language change. In: Briscoe (2002b).

NIYOGI, P. & BERWICK, R. C. (1995). The logical problem of language change. Tech. rep., M.I.T.

NOOTEBOOM, S., WIJNEN, F. & WEERMAN, F., eds. (2002). *Storage and Computation in the Language Faculty*. Studies in Theoretical Psycholinguistics. Dordrecht: Kluwer Academic Publishers.

NOTTEBOHM, F. (1976). Vocal tract and brain: A search for evolutionary bottlenecks. *Annals of the New York Academy of Sciences* **280**, 643–649.

NOWAK, M. A. (2000). Evolutionary biology of language. *Phil Trans R Soc Lond* **355**, 1615–1622.

NOWAK, M. A., KOMAROVA, N. & NIYOGI, P. (2001). Evolution of universal grammar. *Science* **291**, 114–118.

NOWAK, M. A., KOMAROVA, N. L. & NIYOGI, P. (2002). Computational and evolutionary aspects of language. *Nature* **417**, 611–617.

NOWAK, M. A., KRAKAUER, D. & DRESS, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London* **266**, 2131–2136.

NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.

NOWAK, M. A., PLOTKIN, J. B. & JANSEN, V. A. (2000). The evolution of syntactic communication. *Nature* **404**, 495–498.

O'DONNELL, T. (2004). Experimental formal language theory in comparative biology. Tech. rep., Language Evolution & Computation research unit, University of Edinburgh.

O'DONNELL, T. & ZUIDEMA, W. (2004). Mathematical linguistics and language evolution. In: *Proceedings of the fifth Evolution of Language conference*. Leipzig, Germany.

OKASHA, S. (2003). Biological altruism. In: *The Stanford Encyclopedia of Philosophy* (Zalta, E. N., ed.).

OLIPHANT, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior* **7**.

OLIPHANT, M. & BATALI, J. (1996). Learning and the emergence of coordinated communication. *Center for research on language newsletter* **11**, 1–46.

OUDEYER, P.-Y. (2001). Coupled neural maps for the origins of vowel systems. In: *Proceedings of the International Conference on Artificial Neural Networks, LNCS 2130* (G. Dorffner, H. Bischof, K. H., ed.), pp. 1171–1176. Berlin: Springer Verlag.

OUDEYER, P.-Y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In: *Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior* (Hallam, B., Floreano, D., Hallam, J., Hayes, G. & Meyer, J.-A., eds.), pp. 406–416. Cambridge, MA: MIT Press.

OUDEYER, P.-Y. (2003). *L'auto-organisation de la parole*. Ph.D. thesis, University Paris VI.

PARKER, G. A. & MAYNARD SMITH, J. (1990). Optimality theory in evolutionary biology. *Nature* **348**, 27–33.

PAYNE, R. S. & MCVAY, S. (1971). Songs of humpback whales. *Science* **173**, 585–597.

PERRUCHET, P. & REY, A. (2004). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review* In press.

PFEIFER, R. & SCHEIER, C. (1999). *Understanding Intelligence*. Cambridge, MA: Bradford Books/MIT Press.

PIERREHUMBERT, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In: *Frequency effects and the emergence of linguistic structure* (Bybee, J. & Hopper, P., eds.), pp. 137–57. Amsterdam, the Netherlands: John Benjamins.

PINKER, S. (1979). Formal models of language learning. *Cognition* **7**, 217–283.

PINKER, S. (1994). *The language instinct, how the mind creates language*. Harper Perennial.

PINKER, S. & BLOOM, P. (1990). Natural language and natural selection. *Behavioral and brain sciences* **13**, 707–784.

PINKER, S. & JACKENDOFF, R. (2004). The faculty of language: What's special about it. *Cognition*. in press.

PLOTKIN, J. B. & NOWAK, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology* pp. 147–159.

PLOTKIN, J. B. & NOWAK, M. A. (2001). Major transitions in language evolution. *Entropy* **3**, 227–246.

POLLACK, J. B. (1988). Recursive auto-associative memory: Decising compositional distributed representations. In: *Proc. of the Tenth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum.

POLLARD, C. & SAG, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

PREMACK, D. (1971). Some general characteristics of a method for teaching language to organisms that do not ordinarily acquire it. In: *Cognitive Process of Non-Human Primates* (Jarrard, L. E., ed.), pp. 47–82. New York: Academic Press.

PRICE, G. R. (1970). Selection and covariance. *Nature* **227**, 520–521.

PRINCE, A. & SMOLENSKY, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell.

PROVINE, W. (1986). *Sewall Wright and evolutionary biology*. Chicago, IL: University of Chicago Press.

PULLUM, G. K. & SCHOLZ, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review* **19**, 9–50. Special issue: A Review of "The Poverty of Stimulus Argument", edited by Nancy Ritter.

RAMBOW, O. & JOSHI, A. (1994). A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena. In: *Current Issues in Meaning–Text Theory* (Wanner, L., ed.). London, UK: Pinter.

REBY, D., MCCOMB, K., CARGNELUTTI, B., DARWIN, C., FITCH, W. T. & CLUTTON-BROCK, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society, London B* in press.

REDFORD, M. A., CHEN, C. C. & MIIKKULAINEN, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech* **44**, 27–56.

REICH, P. A. (1969). The finiteness of natural language. *Language* **45**, 831–843.

RISSANEN, J. & RISTAD, E. (1994). Language acquisition in the MDL framework. In: *Language Computation* (Ristad, E. S., ed.). Philadelphia: American Mathematical Society.

VAN ROOIJ, R. (2004). Evolution of conventional meaning and conversational principles. *Synthese (Knowledge, Rationality and Action)* **139**, 331–366.

ROSS, J. R. (1967). *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.

ROUGHGARDEN, J. (1979). *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. New York: Macmillan. Reprinted 1987.

RUMELHART, D. & MCCLELLAND, J. (1986). On learning past tenses of English verbs. In: *Parallel Distributed Processing, Vol. 2* (Rumelhart, D. & McClelland, J., eds.), pp. 318–362. Cambridge, MA: MIT Press.

SAFFRAN, J., SENGHAS, A. & TRUESWELL, J. (2001). Language acquisition by children. *Proceedings of the National Academy of Sciences* **98**, 12874–12875.

SAKOE, H. & CHIBA, S. (1978). Dynamic programming optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**, 43–49.

SAVAGE-RUMBAUGH, S. (2000). Paper presented at the evolution of language conference. Paris, France.

SAVAGE-RUMBAUGH, S. & LEWIN, R. (1994). *Kanzi: the ape at the brink of the human mind*. Wiley.

SAVAGE-RUMBAUGH, S., MCDONALD, K., SEVCIK, R. A., HOPKINS, W. D. & RUBERT, E. (1986). Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*). *Journal of Experimental Psychology: General* **115**, 211–235.

SCHOLZ, B. C. & PULLUM, G. K. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review* **19**, 185–224. Special issue: A Review of "The Poverty of Stimulus Argument", edited by Nancy Ritter.

SEIDENBERG, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science* **275**, 1599–1603.

SENGHAS, A., KITA, S. & ÖZYÜREK, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science* **305**, 1779–1782.

SEYFARTH, R., CHENEY, D. & MARLER, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science* **210**, 801–803.

SEYFARTH, R. M. & CHENEY, D. L. (1997). Some general features of vocal development in nonhuman primates. In: *Social influences on vocal development* (Snowdon, C. T. & Hausberger, M., eds.), pp. 249–273. Cambridge, U.K.: Cambridge University Press.

SHANNON, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal* **27**, 379–423 and 623–656.

SHERMAN, P. W. (1977). Nepotism and the evolution of alarm calls. *Science* **197**, 1246–1253.

SHIEBER, S. M. (1985). Evidence against the non-context-freeness of natural language. *Linguistics and Philosophy* **8**.

SHILLCOCK, R., KIRBY, S., MCDONALD, S. & BREW, C. (2004). Exploring systematicity in the mental lexicon. unpublished manuscript, University of Edinburgh.

SIEGELMANN, H. & SONTAG, E. (1991). Neural networks are universal computing devices. Tech. Rep. SYCON–91–08, Rutgers Center for Systems and Control.

SIMON, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics* **69**, 99–118.

SIMON, H. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press. The Karl Taylor Compton lectures.

SKYRMS, B. (1996). *Evolution of the Social Contract*. Cambridge, UK: Cambridge University Press.

SMITH, A. D. M. (2003a). *Evolving Communication through the Inference of Meaning*. Ph.D. thesis, Theoretical and Applied Linguistics, University of Edinburgh.

SMITH, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science* **14**, 65–84.

SMITH, K. (2003b). *The Transmission of Language: models of biological and cultural evolution*. Ph.D. thesis, Theoretical and Applied Linguistics, University of Edinburgh.

SMITH, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology* **228**, 127–142.

SMITH, K. & HURFORD, J. R. (2003). Language evolution in populations: extending the iterated learning model. In: *Advances in Artificial Life (Proceedings of the 7th European Conference on Artificial Life* (Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T. & Ziegler, J., eds.), vol. 417 of *Lecture Notes in Artificial Intelligence*, pp. 507–516. Berlin: Springer-Verlag.

SOLOMONOFF, R. (1960). A new method for discovering the grammars of phrase structure languages. In: *Information Processing*. Unesco, Paris.

STEEDMAN, M. (2000). *The Syntactic Process*. Cambridge, MA: MIT Press/Bradford Books.

STEEDMAN, M. (2002a). Connectionist and symbolic representations of language. In: *Encyclopedia of Cognitive Science*. Nature Publishing Group, Macmillan. (to appear).

STEEDMAN, M. (2002b). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy* **25**, 723–753.

STEEDMAN, M. & BALDRIDGE, J. (2003). Combinatory Categorial Grammar. Unpublished Tutorial Paper, University of Edinburgh, http://www.iccs.informatics.ed.ac.uk/~steedman/papers.html.

STEELS, L. (1995). A self-organizing spatial vocabulary. *Artificial Life* **2**, 319–332.

STEELS, L. (1997). The synthetic modeling of language origins. *Evolution of Communication* **1**, 1–35.

STEELS, L. (1998). Synthesising the origins of language and meaning. In: Hurford *et al.* (1998).

STEELS, L. (2004). Constructivist development of grounded construction grammars. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. San Francisco, CA: Morgan Kaufman.

STEELS, L., KAPLAN, F., MCINTYRE, A. & VAN LOOVEREN, J. (2002). Crucial factors in the origins of word-meaning. In: Wray (2002).

STEELS, L. & OUDEYER, P.-Y. (2000). The cultural evolution of syntactic constraints in phonology. In: *Proceedings of the VIIth Artificial life conference (Alife 7)* (Bedau, M. A., McCaskill, J. S., Packard, N. H. & Rasmussen, S., eds.). Cambridge (MA): MIT Press.

STOLCKE, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley.

STUDDERT-KENNEDY, M. (1983). On learning to speak. *Human Neurobiology* **2**, 191–195.

STUDDERT-KENNEDY, M. (1998). The particulate origins of language generativity: from syllable to gesture. In: Hurford *et al.* (1998).

STUDDERT-KENNEDY, M. (2000). Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function. In: Knight *et al.* (2000).

STUDDERT-KENNEDY, M. (2002). Mirror neurons, vocal imitation and the evolution of particulate speech. In: *Mirror Neurons and the Evolution of the Brain and Language* (Stamenov, M. & Gallese, V., eds.), pp. 207–227. Amsterdam: John Benjamins.

SZATHMÁRY, E. (1993). Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proc. Natl. Acad. Sci. (USA)* **90**, 9916–9920.

SZATHMÁRY, E. (1999). The first replicators. In: *Levels of Selection in Evolution* (Keller, L., ed.), pp. 31–52. Princeton, NJ: Princeton University Press.

TALLERMAN, M., ed. (2004). *Evolutionary Prerequisites for Language*. Oxford, UK: Oxford University Press. (forthcoming).

TALLERMAN, M. (2005). Initial syntax and modern syntax: Did the clause evolve from the syllable? In: *Language Origins: Perspectives on Evolution* (Tallerman, M., ed.), chap. 6. Oxford University Press.

TERRACE, H. S. (1979). *Nim*. New York: Knopf.

THOMPSON, D. W. (1932). *On Growth and Form*. Cambridge, UK: Cambridge University Press. This edition 1942.

TINBERGEN, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie* **20**, 410–433.

TITZE, I. R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice-Hall.

TOMASELLO, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Science* **4**, 156–163.

TOMASELLO, M. (2003). Different origins of symbols and grammar. In: Christiansen & Kirby (2003a), pp. 94–110.

TOMASELLO, M. & BATES, E., eds. (2001). *Language Development: The Essential Readings*. Malden, MA: Blackwell.

TRAPA, P. E. & NOWAK, M. A. (2000). Nash equilibria for an evolutionary language game. *Journal of Mathematical Biology* **41**, 172–188.

TRAUNMÜLLER, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America* **88**.

TRIVERS, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology* **46**, 35–57.

UJHELYI, M. (1996). Is there any intermediate stage between animal communication and language? *Journal of Theoretical Biology* **180**, 71–76.

VOGT, P. (2000). *Lexicon Grounding on Mobile Robots*. Ph.D. thesis, Vrije Universiteit Brussel, AI Lab.

WADDINGTON, C. H. (1939). *An Introduction to Modern Genetics*. London: Allen Unwin.

WAIBEL, A., HANAZAWA, T., HINTON, G., SHIKANO, K. & LANG, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Acoustics Speech and Signal Processing* **37**, 328–339.

WANG, J. (2004). Language evolution and computation bibliography and resources. http://www.isrl.uiuc.edu/amag/langev/.

WEINBERG, W. (1908). Über den nachweis der Vererbung beim Menschen. *Jahresh. Wuertt. Ver. vaterl. Natkd.* **64**, 369–382.

WESTERMANN, G. (2001). A model of perceptual change by domain integration. In: *Proceedings of the 23d Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

WEXLER, K. (1999). Innateness of language. In: *The MIT Encyclopedia of the Cognitive Sciences* (Wilson, R. A. & Keil, F. C., eds.), pp. 408–409. MIT Press.

WEXLER, K. & CULICOVER, P. (1980). *Formal principles of language acquisition*. Cambridge MA: MIT Press.

WEXLER, K. & HAMBURGER, H. (1973). Identifiability of a class of transformational grammars. In: *Approaches to Natural Language* (Hintikka, K. J. J., Moravcsik, J. M. E. & Puppes, P., eds.), pp. 153–156. Dordrecht: Reidel.

WIEHE, T. (1997). Model dependency of error thresholds: The role of fitness functions and contrasts between finite and infinite sites models. *Genetical Research Cambridge* **69**, 127–136.

WILLIAMS, G. C. (1966). *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.

WOLFF, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication* **2**, 57–89.

WOLFRAM, S. (2002). *A New Kind of Science*. Champaign, IL: Wolfram Media.

WOODS, W. A. (1968). Procedural semantics for a question-answering machine. In: *AFIPS Conference Proceedings*. Fall Joint Computer Conference.

WORDEN, R. (1998). The evolution of language from social intelligence. In: Hurford *et al.* (1998).

WRAY, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication* **18**, 47–67.

WRAY, A. (2000). Holistic utterances in protolanguage: The link from primates to humans. In: *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form* (Chris Knight, J. R. H. & Studdert-Kennedy, M., eds.). Cambridge: Cambridge University Press.

WRAY, A., ed. (2002). *The Transition to Language*. Oxford, UK: Oxford University Press.

WRIGHT, S. (1931). Evolution in mendelian populations. *Genetics* **16**, 97–159.

WRIGHT, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: *Proceedings of the Sixth International Congress on Genetics*, pp. 355–366.

YAMAUCHI, H. (2001). The difficulty of the Baldwinian account of linguistic innateness. In: *Advances in Artificial Life (Proceedings 6th European Conference on Artificial Life, Prague)* (Kelemen, J. & Sosík, P., eds.), vol. 2159 of *Lecture Notes in Computer Science*, pp. 391–400. Berlin: Springer.

YANG, C. D. (2000). Internal and external forces in language change. *Language Variation and Change* **12**, 231–250.

VAN ZAANEN, M. & ADRIAANS, P. (2001). Comparing two unsupervised grammar induction systems: Alignment-based learning vs. EMILE. In: *Proceedings of BNAIC 2001* (Kröse, B., de Rijke, M., Schreiber, G. & van Someren, M., eds.).

ZAHAVI, A. (1975). Mate selection - a selection for a handicap. *Journal of Theoretical Biology* **53**, 205–214.

ZAHAVI, A. (1977). The cost of honesty (further remarks on the handicap principle). *Journal of Theoretical Biology* **67**, 603–605.

ZUBERBÜHLER, K. (2002). A syntactic rule in forest monkey communication. *Animal Behaviour* **63**, 293–299.

ZUIDEMA, W. (2000). *Evolution of syntax in groups of agents*. Master's thesis, Utrecht University, Theoretical Biology.

ZUIDEMA, W. (2003a). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.), pp. 51–58. Cambridge, MA: MIT Press.

ZUIDEMA, W. (2003b). Modeling language acquisition, change and variation. In: *Proceedings of the Workshop on Language Evolution and Computation* (Kirby, S., ed.). 15th European Summer School in Logic Language and Information (ESSLLI).

ZUIDEMA, W. (2003c). Optimal communication in a noisy and heterogeneous environment. In: *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)* (Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T. & Ziegler, J., eds.), vol. 2801 of *Lecture Notes in Artificial Intelligence*, pp. 553–563. Berlin: Springer Verlag.

ZUIDEMA, W. & DE BOER, B. (2003). How did we get from there to here in the evolution of language? *Behavioral and Brain Sciences* **26**, 694–695.

ZUIDEMA, W. & HOGEWEG, P. (2000). Selective advantages of syntactic language: a model study. In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (Gleitman & Joshi, eds.), pp. 577–582. Mahwah, NJ: Lawrence Erlbaum Associates.

ZUIDEMA, W. & WESTERMANN, G. (2003). Evolution of an optimal lexicon under constraints from embodiment. *Artificial Life* **9**, 387–402.

APPENDIX A

Wright's Adaptive Topography

Consider the single locus, two alleles model of figure 2.1. Recall the expression for average fitness of the 3 possible genotypes (equation 2.5):

$$\overline{w} = p^2 w_{AA} + 2pqw_{Aa} + q^2 w_{aa} \tag{A.1}$$

Because $p + q = 1$ this expression can be rewritten as:

$$\begin{aligned} \overline{w} &= p^2 w_{AA} + 2p(1-p)w_{Aa} + (1-p)^2 w_{aa} \\ &= p^2 w_{AA} + 2pw_{Aa} - 2p^2 w_{Aa} + w_{aa} - 2pw_{aa} + p^2 w_{aa}. \end{aligned} \tag{A.2}$$

The derivative of $\overline{w}$ with respect to p is now (provided the fitness coefficients are independent of p):

$$\begin{aligned} \frac{\mathrm{d}\overline{w}}{\mathrm{d}p} &= 2pw_{AA} + 2w_{Aa} - 4pw_{Aa} - 2w_{aa} + 2pw_{aa} \\ &= 2\left(pw_{AA} + w_{Aa} - 2pw_{Aa} - w_{aa} + pw_{aa}\right) \\ &= 2\left(pw_{AA} + (1-p)w_{Aa} - pw_{Aa} - (1-p)w_{aa}\right) \\ &= 2\left(pw_{AA} + qw_{Aa} - pw_{Aa} - qw_{aa}\right) \\ &= 2\left(p\left(w_{AA} - w_{Aa}\right) - q\left(w_{aa} - w_{Aa}\right)\right). \end{aligned} \tag{A.3}$$

Now, recall the expression for the change in p (equation (2.6)), which can in a few steps be rewritten as:

$$\begin{aligned} \Delta p &= p' - p \\ &= \frac{p(pw_{AA} + qw_{Aa})}{\overline{w}} - p \\ &= \frac{p(pw_{AA} + qw_{Aa})}{\overline{w}} - \frac{p\overline{w}}{\overline{w}} \\ &= \frac{p}{\overline{w}}(pw_{AA} + qw_{Aa} - \overline{w}). \end{aligned} \tag{A.4}$$

Inserting equation (A.1) into equation (A.4), and rearranging using the fact that $q = 1 - p$, gives:

$$\begin{aligned}
\Delta p &= \frac{p}{\overline{w}}(pw_{AA} + qw_{Aa} - p^2w_{AA} - 2pqw_{Aa} - q^2w_{aa}) \\
&= \frac{p}{\overline{w}}(pw_{AA} - p^2w_{AA} + qw_{Aa} - 2pqw_{Aa} - q^2w_{aa}) \\
&= \frac{p}{\overline{w}}(p(w_{AA} - pw_{AA}) + qw_{Aa} - 2pqw_{Aa} - q^2w_{aa}) \\
&= \frac{p}{\overline{w}}(p(1-p)w_{AA} + qw_{Aa} - 2pqw_{Aa} - q^2w_{aa}) \\
&= \frac{p}{\overline{w}}(pqw_{AA} + qw_{Aa} - 2pqw_{Aa} - q^2w_{aa}) \\
&= \frac{pq}{\overline{w}}(pw_{AA} + w_{Aa} - 2pw_{Aa} - qw_{aa}) \\
&= \frac{pq}{\overline{w}}(pw_{AA} + w_{Aa} - pw_{Aa} - pw_{Aa} - qw_{aa}) \\
&= \frac{pq}{\overline{w}}(pw_{AA} + (1-p)w_{Aa} - pw_{Aa} - qw_{aa}) \\
&= \frac{pq}{\overline{w}}(pw_{AA} + qw_{Aa} - pw_{Aa} - qw_{aa}) \\
&= \frac{pq}{\overline{w}}(p(w_{AA} - w_{Aa}) - q(w_{aa} - w_{Aa})). \qquad \text{(A.5)}
\end{aligned}$$

Equation (A.5) and (A.3) can be combined into equation (2.9), as is explored in the main text.

APPENDIX B

Local Optimisation of a Deterministic Lexicon

Distributed hillclimbing:

For g=0 to I do

$i \leftarrow$ random integer, $0 < i < P$

$j \leftarrow$ random integer, $0 < j < P, j \neq i$

$m \leftarrow$ random integer, $0 < m < M$

$f \leftarrow$ random integer, $0 < f < F$

if g is even do

$w \leftarrow$ quicksuccess-m(S^i, R^j, U, V, m)

$f' \leftarrow S^i[m]$

$S^i[m] \leftarrow f$

$w' \leftarrow$ quicksuccess-m(S^i, R^j, U, V, m)

if $w > w'$ do $S^i[m] \leftarrow f'$

else do

$w \leftarrow$ quicksuccess-f(S^j, R^i, U, V, f)

$m' \leftarrow R^i[f]$

$R^i[f] \leftarrow m$

$w' \leftarrow$ quicksuccess-f(S^j, R^i, U, V, f)

if $w > w'$ do $R^i[f] \leftarrow m'$

quicksuccess-m$(S, R, U, V, m) \leftarrow$

$\sum_{f=0}^{F}$ V[m][R[f]] $\times$ U[S[m]][f]

quicksuccess-f$(S, R, U, V, f) \leftarrow$

$\sum_{m=0}^{M}$ V[m][R[f]] $\times$ U[S[m]][f]

APPENDIX C

Publications

- Willem Zuidema and Paulien Hogeweg, *Selective advantages of syntactic language – a model study* (2000), in: Gleitman and Joshi (eds.), Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, Lawrence Erlbaum Associates, Hillsdale, USA, pp. 577 - 582.
- Willem Zuidema, *Emergent syntax: the unremitting value of computational modeling for understanding the origins of complex language* (2001), in: Kelemen and Sosik (eds.), Advances in Artificial Life (Proceedings 6th European Conference on Artificial Life), Lecture Notes in Computer Science, vol. 2159, Springer Verlag, Berlin, pp. 641-644.
- Willem Zuidema, *The importance of social learning in the evolution of cooperation and communication* (2002), Behavioral and Brain Sciences, vol. 25, issue 2, pp 283-284.
- Joachim De Beule, Joris Van Looveren and Willem Zuidema, *From perception to language: grounding formal syntax in an almost real world* (2002), Proceedings of the Belgium-Netherlands Artificial Intelligence Conference, 21-22 oktober 2002, Leuven, Belgium..
- Willem Zuidema, *Language adaptation helps language acquisition* (2002), in: Bridget Hallam, Dario Floreano, John Hallam, Gillian Hayes and Jean-Arcady Meyer (Eds.), From Animals to Animats 7 (Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior, Edinburgh, August 4-9, 2002), MIT Press, Cambridge, MA, pp. 417-418
- Willem Zuidema, *Optimal Communication in a Noisy and Heterogeneous Environment* (2003), in: W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler, editors, Advances in Artificial Life (Proceedings of the 7th European Conference on Artificial Life), Lecture Notes in Computer Science, vol. 2801, pp. 553-563, Springer Verlag, Berlin.
- Willem Zuidema and Gert Westermann, *Evolution of an Optimal Lexicon under Constraints from Embodiment* (2003), Artificial Life, vol. 9, issue 4, pp 387-402.
- Willem Zuidema and Bart de Boer, *How did we get from there to here in the evolution of language?* (2003). Behavioral and Brain Sciences, vol. 26, issue 6, pp 694–695.
- Nick Barton and Willem Zuidema, *Evolution: The erratic path towards complexity* (2003), Current Biology, vol. 13, issue 16, pp. 649-651.
- Andy Gardner and Willem Zuidema, *Is evolvability involved in the origin of modular variation?* (2003), Evolution, vol. 57, nr. 6, pp 1448-1450.

- Willem Zuidema, *How the poverty of the stimulus solves the poverty of the stimulus* (2003), in: Suzanna Becker, Sebastian Thrun, and Klaus Obermayer (eds.), Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02), MIT Press, Cambridge, MA, pp. 51-58.
- Willem Zuidema (2003), Modeling Language Acquisition, Change and Variation. In: *Proceedings of the Workshop on Language Evolution and Computation* (Kirby, S., ed.). 15th European Summer School in Logic Language and Information (ESSLLI).
- Bart de Boer and Willem Zuidema (2003) Phonemic coding: Optimal communication under noise? In: *Proceedings of the Workshop on Language Evolution and Computation* (Kirby, S., ed.). 15th European Summer School in Logic Language and Information (ESSLLI).
- Tim O'Donnell and Willem Zuidema (2004). Mathematical linguistics and language evolution. In: *Proceedings of the fifth Evolution of Language conference*. Leipzig, Germany.

Selective advantages of syntactic language — a model study

Willem H. Zuidema*
jelle@csl.sony.fr
Theoretical Biology, Utrecht University
Padualaan 8, 3584 CH Utrecht
The Netherlands

Paulien Hogeweg
p.hogeweg@bio.uu.nl
Theoretical Biology, Utrecht University
Padualaan 8, 3584 CH Utrecht
The Netherlands

Abstract

We study a computational model of the evolution of language in groups of agents to evaluate under which circumstances syntax emerges. The fitness in the model depends on the composition of the population. We find that this fact significantly alters the evolutionary dynamics. If scores are attributed to both speaker and hearer, expressive syntax is hard to obtain. If scores are attributed only to the hearer, syntax develops, but agents loose the willingness to speak. Implications and a possible solution of this paradox are discussed.

Introduction

Among the many differences between human language and other animal communication systems, syntax is widely acknowledged to be particularly important. Syntax allows us to combine a finite set of meaningful units into an unbounded set of combinations. It allows us to speak about events happening at other times and places. It allows us to communicate about causal relations, to phrase questions or imperatives, and to share in detail previous experiences. The emergence of syntactic language is therefore considered to be one of the major transitions in evolution (Szathmáry & Maynard-Smith, 1995).

In the traditional view, syntax reconciles the need for high expressiveness with some of the natural boundary conditions on communication such as memory limitations, errors in distinguishing sounds, or bottlenecks in the transmission of language knowledge. However, present-day language fulfills many more functions than exchanging information, including facilitating social relations, individual expression, increase of status, esthetic experience and perhaps internalizing our knowledge of the world. It is unclear in what way such functions are recent side-effects, or play an important role in explaining the origins of language.

Discussions of such issues tend to be very unsatisfactory, because they seem hardly restricted by empirical or theoretical bounds. *Computational modeling* offers a novel approach to these issues, because such models are at least restricted by whether or not the *combination* of assumptions implemented in the model yield the hypothesized outcome: syntactic language. This paper discusses a simple computational model of an evolving group of communicating individuals and studies under which selection pressures expressive, syntactic language arises. Before describing the model architecture and results, we will first briefly discuss the theoretical background and some related work.

Evolution of language

Probably the most well-known speculation on the origins of human language is the paper of Pinker & Bloom (1990). Pinker & Bloom argue that syntax must originate in a process of evolutionary optimization, because "natural selection" is the only explanation for the origins of complex design in nature. The paper brings together a valuable collection of findings, but from a theoretical perspective it is problematic, because it lacks precision and formalization. In its weakest interpretation the central claim is trivial (there is no doubt that only members of the human species can acquire fluency in a human language) and in its strongest interpretation ("evolution has led to genes that explicitly specify a universal rule system for language") the claim is untenable. However, the lack of a more precise aspect to Pinker & Bloom's work, makes it hard to position their ideas between these extremes.

Moreover, Pinker & Bloom's paper is symptomatic for the popular fallacy in linguistics that one can only choose between two explanations: (i) language originates in a genetic evolution, or (ii) language arises as the spontaneous result of general cognitive skills and social structure. We believe that putting these two explanations in opposition, excludes the most interesting part of the story. Spontaneous pattern formation ("self-structuring") needs a mechanism to set the right parameters, and evolution needs a plausible substrate to operate on. Viewing self-structuring as a substrate for evolution (Boerlijst & Hogeweg, 1991a) offers a fresh perspective that allows one to study how evolution, genetic information, learning, development, embodiment and social structures all interact to shape

*Present address: Sony CSL, 6, Rue Amyot, 75005, Paris, France; webpage: www-binf.bio.uu.nl/~jelle

human language. Note that such an interactionist account differs fundamentally from a naive "some parts of language are innate, some are learned" view.

Computational modeling

Recent work that studied such interactions in computational models has produced a wealth of new hypotheses and insights (Hurford, 1989; Hashimoto & Ikegami, 1996; Batali, 1997; Steels, 1997; De Boer, 1999; Kirby, 2000; Nowak & Krakauer, 1999; Hurford, 2000). Such models are *relatively precise* implementations of the underlying set of assumptions, and allow one to evaluate the internal coherence of such a set. Moreover, they are *productive*, in the sense that they often show unexpected behaviors that help to generate new hypotheses and concepts. And although they are necessarily simplified representations, the fact that their behavior can be experimentally evaluated makes it possible to study more complex phenomena than with analytical methods alone. Computational models therefore pre-eminently can make tractable systems with many variables and interactions.

On the issue of the origins of syntax, a number of intriguing mechanisms have been identified using computational modeling techniques. Although very diverse, they all emphasize the fact that syntax greatly increases the number of possible forms in a language. For instance, Batali (1997), Kirby (2000) and Hurford (2000) studied how *cultural evolution* can account for the emergence of syntax. Although they use several different formalisms, the common idea in this work is that the internal knowledge of language (the infinite "I-language") is transmitted culturally (via a finite "E-language") from one agent to another. This "transmission bottleneck" works as a filter, in which syntactic elements of language typically out-compete non-syntactic elements, because the former are inherently used more often.

Nowak & Krakauer (1999) studied a game-theoretic model of language evolution and identify a different mechanism that can account for the emergence of syntax. Using the matrix representations of Hurford (1989), they infer a "linguistic error limit". Given that an individual makes mistakes in distinguishing sounds with a probability that depends on the similarity between those sounds, Nowak & Krakauer calculate a limit on the number of messages an individual can convey. They show mathematically that *word formation* and *syntax* can help overcome such a limit. Moreover, they show that both non-syntactic and syntactic strategies are *evolutionary stable strategies* (i.e. cannot be invaded by other strategies). However, every mixed strategy can be invaded by every mixed strategy that uses *more* syntactic sentences. Thus, the evolutionary process should lead towards grammar.

Hashimoto & Ikegami (1996) showed that syntax can emerge in an *evolving group* of communicating agents. The agents in their model have an internal rewriting grammar, that generates a formal language using lexical or syntactic strategies. Because there is no limit on the number of rules, both strategies could in principal generate all possible strings in the finite domain that was used. However, at the start of the simulations agents are initialized with just one rule in their grammar. Because mutations add rules one at a time, and expressiveness grows much faster with grammar size using a syntactic strategy, syntactic agents outcompete non-syntactic ones.

An important aspect of Hashimoto & Ikegami's model is that fitness is not a fixed measure, but depends on the kind of grammars that are present in the population. This leads to some counterintuitive results. For instance, they find that the most expressive agents are not necessarily the most successful and that a score for *not being recognized* accelerates the evolution of syntax. These observations are the starting point for the model study reported in this paper.

The model

The model reported in this paper is a variant of the model of Hashimoto & Ikegami (1996). Of the many aspects that might be relevant, we study only one particular type of interaction: between evolutionary dynamics and group structure. We therefore ignore all aspects of grammar, except for the fundamental properties of compositionality and recursion. We ignore semantics, by just attributing scores for successful parsing. And we ignore learning, by assuming that agents end up with the same internal grammar, except from some changes that result from mutations in the innate component of language.

In this simplified model we will show that evolution shapes the linguistic environment of agents, but, conversely, that the group structure also shapes the evolutionary process. This interaction guides evolution in unexpected directions, and, depending on the implemented function of language, can both facilitate and hinder the development of syntax.

The model consists of a population of agents with an internal rewriting grammar, which they inherit with some mutations from their parent. The grammars are context free grammars, with nonterminal and terminal symbols from the small alphabets $V_{nt} = \{S, A, B\}$ and $V_{te} = \{0, 1\}$ respectively. As an extra restriction, we don't allow the "S" at the right-hand sides of rules. At the start of most simulations, agents are initialized with a grammar with just one rule: randomly $S \mapsto 0$ or $S \mapsto 1$. Agents have the ability to derive ("speak") and parse ("understand") strings of 0's and 1's of maximum length 6, using the rules from the grammar. Within these constraints the maximum expressiveness is 126. We define compositionality as using the non-terminals A and B, and recursion as using rules that were used before in the same branch of the rewriting tree.

Agents interact in a set-up of "language games". In every game all agents can speak one string and try to recognize the strings produced by other agents. Every generation a number of games is played and scores are attributed for successful communication. In most simulations, we use an explicit "innovation pressure". This pressure is implemented by discounting scores with the number of times a string is already heard before, and corresponds to a semantic need for a rich repertoire of forms. We designed several scoring schemes that reflect hypotheses on the function of language. The most important schemes are labeled "communication" and "perception":

communication corresponds to a selection pressure to optimize the total of exchanged information, such that both the speaker and the hearer benefit from successful communication. This pressure is implemented by a score for recognition and for being recognized.;

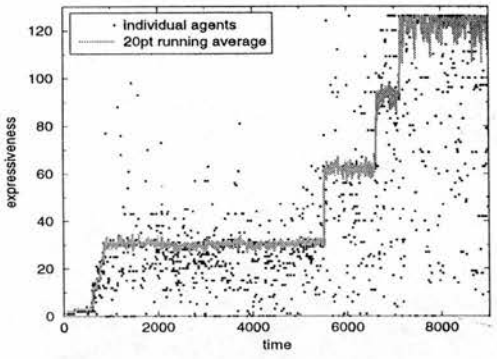
perception corresponds to a selection pressure to optimize the total of information received, in order to make use of the knowledge of others (as if one indirectly shares someone else's *perception*). This pressure is implemented as a score for recognition;

We replace all agents every generation with offspring of the present population. The number of offspring of an agent depends on the total score it has received relative to other agents. Random mutations are applied to the offspring with fixed probabilities for modification of existing rules ("replace"), duplication of a random rule ("add") or deletion of a rule ("delete"). We also implemented a mutation "shift", that swaps a rule with the previous rule in the grammar and occurs with a probability per rule. These mutations correspond to conventions in evolutionary programming and allow for optimizing some of the relevant features of grammars, but otherwise they are more or less arbitrary.

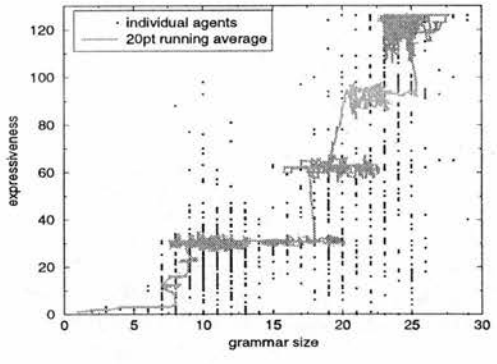
The group effect

Fitness in this model is not a static function of an agent's grammar ("genotype"), but it depends on the grammars of other agents too. The general observation in experiments with the model with many different parameter settings is that this fact strongly influences the evolutionary dynamics (Hashimoto & Ikegami, 1996). The success of an agent's individual language is determined by how well it matches the language of the whole group, rather than by how much information it can encode ("expressiveness"). We call this phenomenon the "group effect".

Figure 1(a) shows an example simulation, with a "communication" scoring scheme and "innovation", that shows clearly some of the mechanisms that play a role. From the initial level of expressiveness of 1, the population evolves within several hundreds of generation to a level of 31. At this point, evolution has developed via selection and random mutations grammars that are redundant and not very structured, and combine several strategies in the rewriting process from the start symbols "S" to a distinct sequence of terminal characters.



(a) Expressiveness over 9000 generations



(b) The same run in a "phase space"

Figure 1: *An example run with very clear epochal evolution. Shown are the running averages and individual agents at every tenth generation. Note that most individual points are hidden under the grey line. (a) During an epoch, expressiveness stays at a fixed level. In fact, in the first stage (E=31) the dominant language stays exactly the same for thousands of generations. Individual agents with higher expressiveness occur, but are not able to survive in the group. (b) Grammars do vary, however, which is possible because of the neutrality in the grammar–language mapping (see text). In the phase space, one can clearly see that grammar size fluctuates during an epoch. All jumps to higher levels take place when grammars are relatively large. Such grammars are clearly larger than necessary and have a neutral tails. Parameters: default "communication" run with innovation pressure (see section "selective advantages")*

For a very long time, from around generation 860 until 5510, the population remains fixed at a level of expressiveness of 31. Analysis of the language reveals that the set of strings of the majority of agents remains unchanged for this whole period. However, frequently agents appear that have a much higher level of expressiveness. This illustrates that (i) the mapping from grammar to language is very non-linear, because a single mutation can make a dramatic change in the size of the language, and (ii) there is a very strong group effect, because agents that have a much higher expressiveness (and thus are "objectively" much better), can nevertheless not persist in the population. The reason is that the languages of these agents differ too much from the language of the group. The agents therefore obtain fewer scores for being recognized and possibly even for recognizing.

Another striking observation in this simulation is that, although the languages remain unchanged for several thousands of generations, the grammars undergo a constant reorganization. This illustrates that the mapping from language to grammar is not only non-linear, but also very redundant.

Figure 1(b) shows a graph of the same simulation in a "phase space" that shows the average grammar size versus the average expressiveness at each generation. As one can clearly see, once a certain level of expressiveness is reached, the evolutionary process "wanders around" for a long time, without significant changes in the expressiveness ("neutrality"). Only when the grammars are relatively large, and thus have many unused, redundant rules, a chance event causes the population to jump to a new level of expressiveness. This chance event is that two agents mutate to the same richer language, and thus can obtain in their mutual communication enough scores to compensate for differing from the group. This mechanism relates to the idea of "neutral networks" — networks of connected points in genotype space that correspond to the same phenotype — that forms a good explanation for the occurrence of "epochs" or "punctuated equilibria" in evolving systems with a fixed fitness function (Van Nimwegen *et al.*, 1999).

Selective advantages

While the "group effect" occurs under all parameter settings of the model, its role can be quite different for each of the scoring schemes and the initial grammars we considered. We observe compositional and recursive grammars only in about half of the parameter combinations we considered. Even if scores are explicitly discounted with the number of times a string is already used before ("innovation pressure"), expressive syntax does not necessarily emerge.

This fact is surprising, because the intuitive expectation is that expressiveness is selectively advantageous. Indeed, with (i) an *explicit* innovation pressure, the average score per agent has it optimum at maximal expressiveness. However, *implicitly* expressiveness influences the scores in other ways as well: (ii) expressive speakers are more likely not to be understood, and (iii) expressive listeners are more likely to understand.

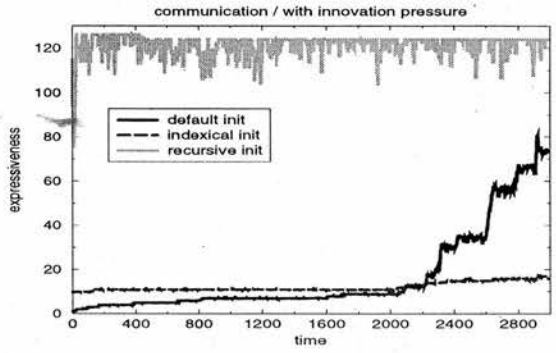


Figure 2: *Communication with innovation pressure for three different types of initial grammars. With a sufficiently large initial lexical grammar, expressive syntax can not develop.*

This leads to an interesting interplay between each of these roles of expressiveness and the group effect. Under communication settings (ii) not being recognized is *disadvantageous*, while (iii) recognition is *advantageous* and in both scoring dimensions similarity to the group's language is important. Under perception settings (ii) not being recognized and (iii) recognition are beneficial, while similarity to the group's language is important for recognition, but *dissimilarity* is better for not being recognized (and thus hindering one's competitors). Moreover, the strength of the group effect depends on the size of the group's language and the variation within the group. In various experiments we obtained the following results:

communication does not lead to highly expressive grammars with the default initial grammar and without the innovation pressure. If the initial grammar is an expressive, recursive grammar, the high level of expressiveness can be maintained. In contrast, with a medium size lexical grammar, grammars remain lexical and expressiveness remains limited.

With an innovation pressure and the default initialization expressive syntax eventually does develop. In this type of runs we observe a stepwise development, with typically long intervals at the same level of expressiveness. Expressive syntactic grammars are reached only after very many generations. With an expressive, recursive initial grammar, the high level of expressiveness can be maintained. With a medium size lexical grammar expressiveness remains limited and no syntax develops (see figure 2).

With "communication" as the function of language,

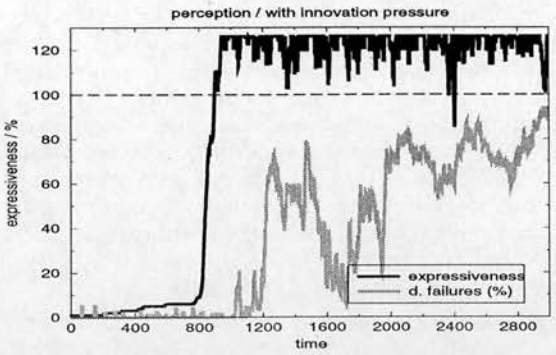


Figure 3: *A typical example of a simulation with "perception" settings, the default initial grammar and an innovation pressure. Shown are the average expressiveness over time, and the percentage of failures in derivation. After around 3000 generations this percentage approaches 100, indicating that very little communication is maintained.*

syntax can thus be maintained if present, but is hard to obtain. If the initial grammar is of sufficient size and of a lexical type, syntax never develops. These results are particularly interesting, as they resemble the situation that is traditionally thought to precede the emergence of grammar: large, lexical protolanguages, with communication benefits for both speaker and hearer.

perception shows rapid growth in expressiveness in most cases considered. With the default initialization and no innovation pressure, expressive syntax develops within a few hundred generations. With the lexical initialization it takes much longer, but the development of syntax was usually observed.

With an innovation pressure and default initial grammars the growth is generally slower than without such an innovation pressure. Infrequently, we even observe runs that remain lexical throughout the simulation. When initialized with an lexical grammar, the runs with innovation pressure show such behavior.

"Perception" thus yields expressive syntax in most cases considered (see figure 3). The benefits of *not being understood* seem to be a strong incentive to develop more expressive language. Interestingly, an innovation pressure makes the development of syntax *less* likely. Apparently, the fact that the hearer benefits from richer input hinders this development.

Paradox

Another striking feature of perception runs is the high number of failures that occur in derivation (see figure 3). Apparently, agents develop grammars that are able to parse a high number of strings, but nevertheless frequently fail in derivation. This is possible because of the asymmetry in parsing (complete bottom-up search of the derivation tree) and derivation (random top-down walk). This possibility was not implemented intentionally. Nevertheless, the evolutionary process discovered it and "actively" exploits it.

This observation points at a important assumption in the model: agents are forced to participate in the language game. A classic altruism problem thus arises: if speaking behavior is beneficial only for an individual's competitors, why would it be retained in evolution? We extended the model with a parameter for probability to speak. Under perception settings this parameter indeed quickly evolves to zero.

Interestingly, these results constitute a paradox: under those circumstances that syntactic expressiveness develops, willingness to speak disappears. Under the circumstances where willingness to speak is retained, syntactic language does not develop. We studied a possible solution for this paradox in a model where agents are localized on a 2D grid and interact only with their immediate neighbors. Such spatial models are known to naturally yield altruism, because spatial patterns make multilevel evolution possible and kin selection more likely (Boerlijst & Hogeweg, 1991b).

The willingness to speak can be retained in the spatial model with perception settings. The parameter that determines the probability of an agent to speak at its turn in the language game, is initialized at 0.1. As one can see in the example of figure 4, the average value rapidly evolves to a high value close to the maximum. Spatial patterns are responsible for this selection pressure towards altruistic behavior. If one destroys the spatial patterns, also the willingness to speak disappears.

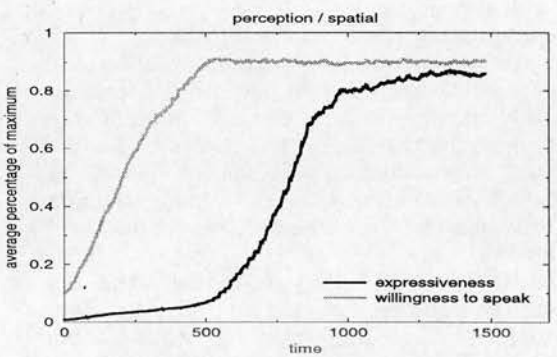


Figure 4: *Perception in space. Shown is the average fraction of the maximum of expressiveness (maximum is 126) and willingness to speak (maximum is 1). Parameters are: initial population size = 2000, number of games per generation = 1, maximum string length = 6, minimum number of understanders = 0, madd = 0.1, mrep = 0.01, mdel = 0.01, maximum number of parsing steps = 500, maximum number of derivation steps = 60, self-interaction not allowed, discount factor 1.0, scores proportional to string length*

Discussion

Some of the striking differences in the results of different scoring schemes can be better understood by looking at a very simple game-theoretic model, where there are just two agents and two levels of expressiveness. If we work out the language games that take place in such a set-up, we find that both the low/low and the high/high situations are equilibria in the communication case, but in the case of perception only the high/high situation is an equilibrium. These results qualitatively corresponds to the results we obtained in the simulations.

The essential observation here is that, although homogeneous high expressiveness is the "best" solution, *unilateral* high expressiveness under communication setting is in fact disadvantageous. It seems a promising approach to extend this game-theoretic analysis to a more general case, with more levels of expressiveness and more interacting agents. However, many aspects of the model behavior depend on the non-linear mapping between grammar and language and can not easily be captured in such an analysis.

Conclusions

Traditionally the origins of language are thought to be explained as either the spontaneous result of human cognitive abilities and social interactions, or the result of an evolution of our innate language capacity. This model study shows an example system where both social interaction and evolutionary updating play a role. Not because one part of language can be explained by "nurture" and another part by "nature", but because they fundamentally interact: social interactions shape the evolutionary process and vice versa.

Also, traditionally language and the evolution of language are studied in terms of how much information about the outside world can be transmitted. Our results suggest that this might not always be the most interesting way of looking at language, because language can have its own dynamics within a group that is quite independent from how well it represents the outside world.

Moreover, this model study shows results that deviate from the traditional picture that lexical protolanguages became larger and larger until syntax became necessary. If communication if beneficial for both speaker and hearer and the population uses an extensive lexical language, syntax does not develop. If the traditional picture holds, the question arises which mechanisms are responsible for the differences.

Finally, spatial patterns have not played much of a role in speculations about the origins of language. Results from this study suggest that such spatial patterns can be relevant. The fact that present-day language shows obvious spatial patterns indicates that a global approximation perhaps excludes important mechanisms.

Many open questions remain. For instance, under perception settings there is an *indirect benefit* of speaking that leads to high values of the willingness to speak. Why then, does this indirect benefit not result in the same disadvantage of unilateral high expressiveness that we observe under communication settings? Such intriguing issues are left for future work.

Acknowledgements

Many thanks to Ludo Pagie for help with programming C++, and Michael Moortgat, Onno Zoeter, and all members of the Theoretical Biology group for technical and moral support and stimulating discussions. WHZ thanks Sony CSL-Paris for allowing him the time to write and present this paper.

References

BATALI, J. (1997). Computational simulations of the emergence of grammar. In: *Approaches to the evolution of language* (Hurford, J. et al., eds.). Cambridge University Press.

BOERLIJST, M. C. & HOGEWEG, P. (1991a). Self-structuring and selection. In: *Artificial Life II* (Langton, C. et al. eds.), 255–276.

BOERLIJST, M. C. & HOGEWEG, P. (1991b). Spiral wave structure in pre-biotic evolution. *Physica D* **48**, 17–28.

DE BOER, B. (1999). *Self-Organisation in Vowel Systems*. Ph.D. thesis, Vrije Universiteit Brussel AI-lab.

HASHIMOTO, T. & IKEGAMI, T. (1996). The emergence of a net-grammar in communicating agents. *BioSystems* **38**, 1–14.

HURFORD, J. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222.

HURFORD, J. R. (2000). Social transmission favours linguistic generalization. In: *The evolutionary emergence of language* (Knight, C. et al., eds.). C.U.P.

KIRBY, S. (2000). Syntax without natural selection. In: *The evolutionary emergence of language* (Knight, C. et al., eds.). C.U.P.

NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.

PINKER, S. & BLOOM, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*

STEELS, L. (1997). Synthesising the origins of language and meaning. In: *Approaches to the evolution of language* (Hurford, J. et al., eds.). C.U.P.

SZATHMÁRY, E. & MAYNARD-SMITH, J. (1995). The major evolutionary transitions. *Nature* **374**, 227–232.

VAN NIMWEGEN, E., CRUTCHFIELD, J. & HUYNEN, M. (1999). Neutral evolution of mutational robustness. *Proc. Nat. Acad. Sci. USA* **96**, 9716–9720.

Emergent syntax: the unremitting value of computational modeling for understanding the origins of complex language

Willem H. Zuidema

Artificial Intelligence Laboratory
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
jelle@arti.vub.ac.be
http://arti.vub.ac.be/~jelle

Abstract. In this paper we explore the similarities between a mathematical model of language evolution and several A-life simulations. We argue that the mathematical model makes some problematic simplifications, but that a combination with computational models can help to adapt and extend existing language evolution scenario's.

1 Introduction

The debate on the origins of language has been dominated by "verbal" theories, both in scientific publications (see e.g. [3]) and in popular, best-selling books (e.g. [1]). Recently also mathematical models of the evolution of language, especially those of Martin Nowak et al., have received much attention (e.g. [6]). These models are sometimes seen as a validation of the earlier verbal theories. Steven Pinker, e.g., writes in the accompanying news story of [7] that the paper shows *"the evolvability of [one of] the most striking features of language"*, i.e. its compositionality.

Although we appreciate the major contributions in these books and papers, we still observe many shortcomings in the proposed theories. Both the verbal and the mathematical accounts tend to overlook many crucial details. Verbal theories often underestimate the intricacies of the evolutionary dynamics and take "evolution" too much as a general problem solver. The mathematical models often make crucial simplifications that are linguistically poorly motivated. In particular, both types of theories have shown little appreciation for the importance of the "frequency dependency" of language evolution and the role of selforganization there-in.

A-life models, on the contrary, have shed light on both the dynamics of language evolution and the explanatory role of selforganization. However, A-life models are too often studied as relatively isolated cases, and too seldomly systematically compared with each other and with mathematical models (the review papers [8, 4] are exceptions, although they unfortunately do not discuss

mathematical models). In this paper we explore the similarities between a recently published mathematical model [6], our own A-life simulations [9] and the model of Kirby [5]. We believe that such an approach can eventually both avoid the problematic simplifications of mathematical models, and the *ad hoc-ness* of many A-life models. In the conference presentation we will also discuss some shortcomings of "verbal" theories as revealed by A-life models.

2 The mathematical model

Nowak et al. use in [6] an elegant formalism that is in line with our view that one should study both the cultural dynamics of language and the evolutionary dynamics that operate on the parameters of the cultural process. We will discuss here only the model for cultural dynamics.

Nowak et al. assume that there is a finite number of states (grammar types) that an individual can be in. Further, they assume that newcomers (infants) learn their grammar from the population, where more successful grammars have a higher probability to be learned and mistakes are made in learning. The system can now be described in terms of the changes in the relative frequencies x_i of each grammar type i in the population:

$$\dot{x_i} = \sum_{j=0}^{N} x_j f_j Q_{ji} - \phi x_i \tag{1}$$

In this differential equation, f_i is the *relative fitness* (quality) of grammars of type i and equals $f_i = \sum_j x_j F_{ij}$, where F_{ij} is the expected communicative success from an interaction between an individual of type i and an individual of type j. The relative fitness f of a grammar thus depends on the frequencies of all grammar types, hence it is *frequency dependent*. The proper way to choose F depends on the characteristics of *language use* (production and interpretation).

Q_{ij} is the probability that a child learning from a parent of type i, will end up with grammar of type j. The probability that the child ends up with the same grammar, Q_{ii}, is defined as q, the copying fidelity. The proper way to choose Q depends on the characteristics of *language acquisition* (learning and development). (ϕ is the average fitness in the population and equals $\phi = \sum_i x_i f_i$. This term is needed to keep the sum of all fractions at 1).

The main result that Nowak et al. obtain is a "coherence threshold": they show mathematically that there is a minimum value for q to keep coherence in the population. If q is lower than this value, all possible grammar types are equally frequent in the population and the communicative success in minimal. If q is higher than this value, one grammar type is dominant; the communicative success is much higher than before and reaches 100% if $q = 1$. Further, Nowak et al. derive an upper and a lower bound on the number of sample sentences that a child needs to acquire its parents' language with the required fidelity q.

3 A-life models

We argue that computational models that we [9] and others [2] have studied fit the general format of equation 1 well, but differ significantly in the particular choices for the representation of language use and language acquisition, i.e. the functions F and Q. In the limited space that is available here we will only shortly mention two examples of interesting, qualitative differences that these choices bring.

First, for sake of simplicity Nowak et al. assume that all grammars are *equally expressive*, and are all *equally similar* to each other. This has the unrealistic consequence that the benefits of interacting with another individual (F) are either maximal or minimal. We studied a computational model [9] were we used context-free grammars to represent the linguistic abilities of agents. This formalism can represent "languages" of many different types and levels of expressiveness. In that study, we did not model learning explicitly, but in stead assumed (as in equation 1) that children end up with a slightly different grammar than their parents.

One of the surprising findings was that once a certain type of language was established in the population, the language kept changing but remained of the same type. The language types formed "self-enforcing regimes", because the language present at time t determines which agents will be successful and reproduce to the next generation, and therefore indirectly determine the language at time $t + 1$. We found three such regimes: (i) idiosyncratic, non-syntactic languages, (ii) compositional languages and (iii) recursive languages. In a population where a rich but idiosyncratic language is established, syntax could not emerge. This phenomenon is important for understanding the consequences of the frequency dependency of language evolution, but is excluded in the simplifications of the mathematical model.

Second, Nowak et al. consider two extreme possibilities for the learning algorithm, and claim to have found a lower and a upper bound on the number of training samples that a learning algorithm needs to reach the coherence threshold. However, in their analysis they have not taken into account that the choice of the grammar that a child has to learn is biased by how well previous generations have been able to learn and maintain it.

In a follow-up of the study above, we have implemented a variant of the "iterated learning model" of Kirby [5], in which agents are endowed with a language-acquisition algorithm to learn the context-free grammars. Kirby found that in the process of iterated cultural transmission the language adapts itself to be better learnable by individual agents. Concretely, this means that the language becomes compositional (syntactic) and that agents are more successful in learning it than would be expected a priori. We replicated this finding, and can show that agents in fact need less training samples than Nowak et al. calculate as a lower bound for maintaining a stable language in the population. The reason is that not only do individuals evolve to be better at language-learning, but also do languages evolve to be better learned [1]. Again, this phenomenon is

important for our understanding of the origins of language, but excluded in the simplifications of the mathematical model.

4 Conclusions

Research on the evolution of language faces two aspects of language that are particularly important: (i) it is transmitted, at least in part, culturally, and learned by one individual from the other; (ii) it is a group phenomenon, that occurs only between individuals and has no apparent value for an individual in isolation. These aspects make that the fitness of individual is not a function of its language acquisition system alone, but is dependent on the cultural dynamics and the composition of the group it is in as well. This observation brings *restrictions* and *opportunities* for language evolution scenario's that are deemed to be overlooked in both verbal and mathematical theorizing. We conclude that A-life models can help to evaluate the validity of these scenarios and help to adapt them, while at the same time mathematical models can help to compare computational models and to identify common themes between them.

Acknowledgements Part of the work reported here has been done in close collaboration with Paulien Hogeweg. I thank her and other members of the Theoretical Biology group in Utrecht, the Netherlands, and members of the AI-Laboratory in Brussels, Belgium, for many helpful discussions.

References

1. Terrence Deacon. *Symbolic species, the co-evolution of language and the human brain.* The Penguin Press, 1997.
2. Takashi Hashimoto and Takashi Ikegami. The emergence of a net-grammar in communicating agents. *BioSystems*, 38:1–14, 1996.
3. J. Hurford, M. Studdert-Kennedy, and C. Knight, editors. *Approaches to the evolution of language.* Cambridge University Press, 1998.
4. James R. Hurford. Expression / induction models of language. In Ted Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models.* Cambridge University Press, 2000.
5. Simon Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, J. Hurford, and M. Studdert-Kennedy, editors, *The Evolutionary Emergence of Language: Social function and the origins of linguistic form.* Cambridge University Press, 2000.
6. Martin A. Nowak, Natalia Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291:114–118, 2001.
7. Martin A. Nowak, Joshua B. Plotkin, and Vincent A.A. Jansen. The evolution of syntactic communication. *Nature*, 404:495–498, 2000.
8. Luc Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1:1–35, 1997.
9. Willem H. Zuidema and Paulien Hogeweg. Selective advantages of syntactic language: a model study. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 577–582. Lawrence Erlbaum Associates, 2000.

freshing to read an article unabashedly stressing the role of the environment in prosocial behaviour. That *in principle* one can imagine behavioural reinforcement as explaining virtually all of the variance in altruistic behaviour, is certainly possible: In the limit, we could have the case where evolution has been evolution for a purely environmentally plastic brain (e.g., Quartz & Sejnowski 1997). One can make a powerful case for some endogeneity of prosocial behavior (Zizzo, in press). Yet, although one can make an argument for the *partial* endogeneity of interdependent preferences, it is not obvious how the evidence is incompatible with a partial role of genetic inheritance in explaining behavioral variance, nor why this should not be considered as the most natural interpretation (e.g., Rushton et al. 1986).

Rachlin's evidence suggests that (1) conditional cooperation is important, and (2) it is subject to framing effects (i.e., whether you believe you are playing against a computer or otherwise). Concerning point 2, Rachlin is right to suggest that framing effects are pervasive (e.g., Cookson 2000). He claims that they are due to different frequencies of reinforcement; this may very well be true, but it is no more than a conjecture, and so it is unclear how it can be used as a proof of Rachlin's theory of altruism relative to other theories. Concerning point 1, reinforcement over an act is not identical to reinforcement over a pattern of acts, and to prove the latter, Rachlin would really need to discuss the evidence on knowledge transfer from one game to a different game, to see whether reinforcement in one situation translates into reinforcement in another situation.

In favour of Rachlin's thesis, there are contexts where this is the case, at least in the short run implied by the laboratory settings (e.g., Guth et al. 1998). In some current experimental work, I have subjects first play a set of games that change to the degree in which the subjects are cooperative or close to zero-sum, and then they play a set of new, never-before-encountered games (with different players, eliminating repeated game effects). When the first set of games is more cooperative, behaviour in the second set of games is also more cooperative. While not all the evidence can be reconciled with a simple reinforcement learning account, it is what Rachlin's theory needs.

A deeper problem is whether the pattern of acts that is reinforced is what Rachlin claims it to be ("altruism") or something entirely different. There are many possible preferences that would be able to explain why cooperation in the finitely repeated Prisoner's Dilemma (PD) is conditional on an expectation of cooperation from the other player. Preferences, as we economists use them, are a behavioral concept: They are preferences as revealed in behaviour and so are closely related to Rachlin's patterns of acts. They include utility functions with two elements, one based on material gain and the other on a payoff transformation component implying inequality aversion (Fehr & Schmidt 1999), reciprocity (Falk & Fischbacher 1998), trust responsiveness (Bacharach et al. 2001), pure or impure altruism (Palfrey & Prisbrey 1997), or perceived fairness (Konow 2000). They will all lead to different predictions depending on whether subjects believe they are playing against a computer, or if they believe they are playing with a human being, because you will be fairness-sensitive towards a human being, but not against a computer. Even if the agents are, and remain, purely self-interested, they may find it optimal to cooperate because of the repeated nature of the game, whether with humans (Kreps et al. 1982) or (in different ways) with computers (because subjects can try to "crack the system" of how to make the most money). Otherwise, for a *wide* range of payoff transformations, with a modicum of rationality, the PD becomes a different game where mutual cooperation is a possible equilibrium, and the greater the expectation is of cooperation from the co-player, the greater will be the expected payoff for cooperating and hence the likelihood of cooperation. Therefore, the interpretation of cooperation in the finitely repeated PD is likely to be difficult. This matters, because the preferences that subjects have or acquire may make very different quantitative predictions in many different game settings (e.g., for other trust games; see Bacharach et al. 2001). This is why experimental economists have been focusing on a variety of different games to assess what preferences subjects have (e.g., Charness & Rabin 2000; Zizzo 2000a): The PD paradigm is simply not discriminative enough.

Rachlin's section 4 definition of altruism appears based on the intrinsic value of an act that is beneficial to a group: This would correspond to what economists would label "impure altruism" or "warm glow," albeit further specified with relation to a group. Unfortunately, there is no specific reason to believe that this *is* the pattern of acts that gets reinforced rather than, say, others with greater predictive power such as inequality aversion (e.g., Fehr & Schmidt 1999; Zizzo 2000b). Perhaps his theory can be rescued by making it more general, but this may be at the cost of virtual unfalsifiability. If Rachlin wants to convince nonbehavioural psychologists, he might need to show how his theory is better than alternative theories that make precise quantitative predictions, and how it can then receive unequivocal support or falsification in the laboratory. Nevertheless, he is right in stressing the role of behavioural reinforcement, and behavioural psychologists like him can bring useful new perspectives to our understanding of prosocial behavior. In particular, framing effects are real, and none of the models I mentioned can really explain them except in specific cases or with auxiliary or unmodelled hypotheses. Zizzo (2000b) tried to fill the modelling gap among reinforcement, framing effects, and preferences using neural network agents learning to play "altruistically" or "enviously" in new games, but this work is very preliminary and tentative.

The importance of social learning in the evolution of cooperation and communication

Willem Zuidema

Language Evolution and Computation Research Unit, Theoretical and Applied Linguistics, University of Edinburgh, Edinburgh EH8 9LL, United Kingdom. **jelle@ling.ed.ac.uk** **http://www.ling.ed.ac.uk/~jelle**

Abstract: The new emphasis that Rachlin gives to social learning is welcome, because its role in the emergence of altruism and communication is often underestimated. However, Rachlin's account is underspecified and therefore not satisfactory. I argue that recent computational models of the evolution of language show an alternative approach and present an appealing perspective on the evolution and acquisition of a complex, altruistic behavior like syntactic language.

Rachlin calls attention to the role of social learning in the emergence of altruistic behavior in humans. This shift of emphasis in thinking about altruism has intriguing consequences. Acknowledging the important role of learning leads one to ask at least three new and challenging questions: (1) about the exact mechanisms by which altruistic behavior emerges in learning and development; (2) about the ways in which the existence of learning mechanisms has changed the evolutionary process; and, vice versa, (3) about the ways in which evolution has shaped the learning mechanisms that lead to altruism. We can no longer – as is common in traditional game-theory – ignore the intricate mapping between genotypes (the genes) and phenotypes (the behaviors) and the strong dependence of this mapping on the individual's (cultural) environment.

Rachlin's article is a welcome effort to underline this point, but I think his explanation for the emergence of altruistic behavior in humans suffers from *underspecification:* Some crucial concepts are too loosely defined to make it possible to really agree or disagree with his analysis. I will discuss Rachlin's answers to the previous questions from this perspective and then try to show that some recent computational models in the related field of the evolution of communication offer a more precise account of the evolution of altruistic behavior.

Rachlin's answer to the first question is a mechanism similar to

self-control. Humans discover that choosing for a whole pattern of altruistic activities is in the end more rewarding than repeating alternative, selfish activities, even though the latter offer more short-term benefits. The problem with this account is that it is unclear what constitutes a "pattern." Without a theory on how individuals represent and acquire this knowledge, we can never identify the different strategies that individuals can choose from.

A related problem arises for Rachlin's implicit answer to the second question. Rachlin gives the example of a woman who puts her life in danger to rescue someone else's child. His explanation of her brave behavior rests on the crucial assumption that the woman at some point in her development had to choose between life-long altruism or life-long selfishness. If there are only these two choices, and if the first choice is indeed more profitable in the long run, natural selection of course favors the tendency to choose it. However, Rachlin gives no arguments why the choice would be so constrained. I find it difficult to accept that with all the subtle influences that genes have on our behavior, selectively avoiding life-threatening situations was not a possibility.

Rachlin's implicit answer to the third question is no solution to that objection. Essentially, he explains the evolution of altruistic behavior by claiming that it is not really altruistic after all. Altruistic – at least in its traditional sense in evolutionary game theory (Maynard Smith 1982) – are those behavioral strategies that benefit others, but harm the individual that employs them *even though less harmful strategies are available.* A game-theoretic analysis of the evolution of alarm calls in certain bird species (Maynard Smith 1982) therefore emphasizes evidence that the calls really are harmful and that other strategies are really available. In contrast, the altruistic strategy in Rachlin's scenario is in the long run advantageous, and better alternatives are not available; it is thus not *really* altruistic in the traditional sense.

Rachlin acknowledges this, but he does not mention that the analogy between his explanation and group selection therefore breaks down. Group selection, like kin selection, is a mechanism that is capable of explaining real altruism. The decrease in the fitness for the *individual* is explained by assuming a higher or lower level of selection, that is, that of the *group* or that of the *gene.* Therefore, the fitness of a worker bee that does not produce any offspring really is low (it is zero by definition), but the fitness of the whole colony or the fitness of the genes that cause her sterility is high. The empirical validity of these explanations remains controversial, although their explanatory power is appealing.

Researchers in the related field of language evolution have already explored many aspects of the interactions between learning and evolution. Language is a complex behavior that is, at least in some cases, used for altruistic purposes (of course, sometimes selfish motives like intimidation, manipulation, and encryption can also play a role). The *population* as a whole benefits from the altruistic use of language, as it does from other altruistic behaviors. In particular, the population benefits from using syntactic language (Pinker & Bloom 1990), but it is not trivial to explain how an *individual* that uses syntax can be successful in a nonsyntactic population.

By using a methodology of computational modeling that avoids the underspecification of Rachlin's arguments, researchers in this field have shed some new light on how this behavior has emerged (Hurford 2002; Steels 1999). For example, these models have shown that when individuals learn language from each other with rather generic learning mechanisms, a rudimentary syntax can emerge without any genetic change (Batali 1998; Kirby 2000). The learning algorithms, for example, the recurrent neural network model in Batali (1998), provide – although far from finally – a fully *specified* candidate answer to the first question we posed previously.

Similarily, in recent work I have explored some provisional answers to the second and third questions. In Zuidema (2003, forthcoming) I explore the consequences of the fact that language itself can, in the process of learners learning from learners, adapt to be more learnable (Kirby 2000). As it turns out, this *cultural* process facilitates the *evolutionary process.* Evolutionary optimization becomes possible, because the cultural learning process fulfills the preconditions for a coherent language in the population. Moreover, the model also shows that much less of the "knowledge of language" needs to be innately specified than is sometimes assumed. Cultural learning thus lifts some of the burden of genetic evolution to explain characteristics of language. Alternatively, Zuidema and Hogeweg (2000) present results of a spatial model of language evolution. These results show that syntax can be selected for through a combined effect of kin selection and group selection.

These answers are far from final, but I believe that such well-defined models present an appealing perspective on *how* cultural learning can lead to the successful acquisition and creation of a complex, altruistic behavior like syntactic language, and *why* the learning mechanisms operate the way they do.

ACKNOWLEDGMENT

The work was funded by a Concerted Research Action fund (COA) of the Flemish Government and the Vrije Universiteit Brussel.

Author's Response

Altruism is a form of self-control

Howard Rachlin

Psychology Department, State University of New York, Stony Brook, NY 11794-2500. **howard.rachlin@sunysb.edu**

Abstract: Some commentators have argued that all particular altruistic acts are directly caused by or reinforced by an internal emotional state. Others argue that rewards obtained by one person might reinforce another person's altruistic act. Yet others argue that all altruistic acts are reinforced by social reciprocation. There are logical and empirical problems with all of these conceptions. The best explanation of altruistic acts is that – though they are themselves not reinforced (either immediately, or delayed, or conditionally, or internally) – they are, like self-controlled acts, part of a pattern of overt behavior that is either extrinsically reinforced or intrinsically reinforcing.

The commentaries demonstrate the enormous variety of approaches that may be taken to explain altruism. Though these approaches each afford a different perspective on the target article, I have attempted to classify them under a few general and overlapping headings. I will discuss each heading in turn, referring to specific commentaries as I go. Although all of the commentaries are thoughtful and deserve thorough discussion, it is not possible in this limited space to answer each commentator in detail. Instead, I have tried to highlight crucial points and respond to common criticisms.

R1. Teleological behaviorism, cognition, and neuroscience

Gray & Braver draw implications from the behavioral correspondence of self-control and altruism for both cognition and neuroscience. Their suggested empirical tests are certainly important and worth doing. But I do not believe that you can crucially test a behavioral model, or even a purely cognitive model, with neurophysiological measurements.

From perception to language: grounding formal syntax in an almost real world

Joachim De Beule Joris Van Looveren Willem Zuidema

Artificial Intelligence Laboratory
Vrije Universiteit Brussel
Pleinlaan 2, B-1050, Brussel, Belgium
{joachim, joris, jelle}@arti.vub.ac.be

Abstract

We present a model that explicitly connects low-level perception and categorization, hierarchical meaning construction and syntactic language. The model shows a solution to the "symbol grounding problem" [7]: the meaning of the symbolic system—logical symbols and syntactic rules—is grounded in its relation with a simplified but realistic world. We discuss the different components of this collaborative effort: (i) a realistic simulation of Newtonian dynamics of objects in a 2D plane; (ii) schema-based event-perception and categorization; (iii) a semantics based on predicate logic; and (iv) a categorial grammar for the production and interpretation of language. The integration of the different components poses on the one hand novel and important constraints; on the other hand, it allows for experiments that help to identify the relations between the different levels.

1 Introduction

In recent years artificial life researchers have begun to study the origins and nature of human language. Such studies have concentrated on the emergence of a lexicon of arbitrary form-meaning mappings, through evolutionary optimization [11, 15] or through coupled learning in populations of agents [8, 19, 16]. With the same approach, interesting results have also been obtained on the emergence of sound systems [4] and syntax [9, 2]. These studies focused on issues that were largely ignored in more traditional approaches, in particular the question: how did human languages become the way they are? Moreover, they put an emphasis on studying "complete agents" that not only are able to *interpret* language (the focus of most work in linguistics), but also to *produce* and *acquire* language, and use it to perform a *task* in an *environment*. This emphasis shift brings new criteria on what makes useful representations, formalisms and models, and it brings new challenges.

Harnad [7] defines one of these challenges as the "symbol grounding problem": *how is symbol meaning to be grounded in something other than just more meaningless symbols?* Harnad argues that it is cognitive theory's burden to explain how "human beings (or any other devices) [...] can (1) discriminate, (2) manipulate, (3) identify and (4) describe the objects, events and states of affairs in the world

they live in", and "can also (5) 'produce descriptions' and (6) 'respond to descriptions' of those objects, events and states of affairs." [7]. We present here a system that is, for a simplified world, capable of doing all these things.

But our ambitions go further. If we are to seriously explore the functional and semantic constraints on the (1) use, (2) acquisition and (3) evolution of language, we need a sophisticated model of meaning that is grounded in interactions with the world. Most existing models of the evolution of syntactic language [e.g. 2, 9] presuppose the existence of a set of (extremely simple) meanings. In this article we describe a system that was designed to investigate the acquisition and evolution of a language that is grounded in a rich interaction with the world. The system has similarities with other attemps to build integrated systems, most notably SHRDLU [22] and the Talking Heads (TH) experiment [20]. The most important differences with these systems are that the present system is much more adaptive (the knowledge of the world is not pre-programmed as in SHRDLU) and (unlike the Talking Heads) can deal with a dynamic world and grammatical language.

2 Simulating Newtonian dynamics

As a first step towards our new system, we built a virtual and simulated world for our agents to live in. We need a world complex enough to allow for hidden states, time, causation, etc. A simple yet realistic model of part of the real world seems a good candidate. Therefore we chose to implement a blocks world that is two-dimensional and only consists of rigid polygonal bodies, but where the bodies actually behave as prescribed by the laws of Newtonian physics, including rotation, static and dynamic friction, gravity, etc.

There is a vast amount of literature on how to implement rigid body simulations, e.g. [13, 1, 12]. What is important for our purposes is that it is possible to simulate real undeformable world objects very realistically. Although we will not elaborate about technical and mathematical details of rigid body dynamics, we mention that colliding contacts are handled by impulse forces including friction and energy dissipation (see [6, 14] and [3] for a thorough analysis of the subject), while resting and sliding contacts are handled by action/reaction forces as described in [1], slightly modified to allow for fast friction force calculation. Using this scheme, the behavior of simulated rigid bodies is very realistic.

At every simulation step we can ask the system to provide us with information about its current state. For example we can get information about the position of blocks in the simulation. This information can be used to supply an agent with input or observations. Figure 1 gives a sequence of frames for a simple simulation of a ball bouncing down some stairs and colliding with some domino bricks.

3 Conceptualization

The simulation described in the previous section provides input for an agent. It defines the world of interaction for an agent; the world the agent should observe, reason about and act upon. The first question at this point is what type of data

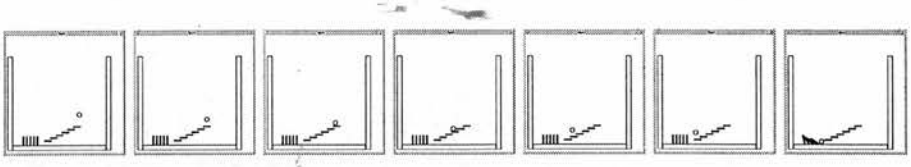


Figure 1: A ball, initially rotating anti-clockwise, bounces down the stairs.

from the simulation should be given to an agent. For example, we could provide an agent with all pixel values of a simulation window. The other extreme would be to give the agent access to the entire state of the simulation (positions, impulses, contacts, etc.) This would not be consistent with our aim and philosophy of *grounding* both the origin and meaning of an agent's concepts, for then, the world would again be part of the agent. Therefore, we transform (not copy) part (not everything) of the state of the simulation to a new set of observation variable-value pairs. This *raw data* represents the equivalent of data an animal receives through its receptors. Of course, it *is* not equivalent, it only *represents* perceptual data as one would define it in a real animal. However, the agent does not get information from the simulation that could not be provided by applying segmentation and other standard image processing techniques on camera images [as in the TH, 18].

The next question is: what should the event detection or *concept formation* system do with this raw data? It should be able to detect meaningful and thus useful concepts with respect to the agent's task (e.g. a language or discrimination game, prediction, "moving all red blocks to the left"). In addition, it should be able to create new detector channels, new concepts triggered by other primitive or previously created event detectors. For example, assume an agent is at some point able to measure, through observation, positions of objects in the world and it would be useful for the agent to create a detector for approaching objects. A new approach detector could be built, looking for pairs of objects of which the positions are getting closer. Note that, because newly created detectors in their turn become building blocks for other event detectors, arbitrarily abstract concepts could come to existence. As explained in [7], these will still be grounded. Finally, it should be possible to attach actions and predictions to the occurrence of an event.

3.1 Implementation

Our implementation consists of *item* and *template* data-structures, together with a system for processing them. An item has a set of features, every feature having a name, a value and a history (see later). Observation of the simulation results in a set of object items with features for position, color, etc. Other items representing more abstract or new channels should be *created.* This is the task of templates. Templates are detectors for various things: simple object items, but also for configurations of objects (e.g. "tower") or events (e.g. approaches or touches). Basic templates consist of an activation slot and an action slot. The activation slot is a set of conditions on items the template needs to get activated. For example, the approach template mentioned above, could have an activation slot

```
(and (has-feature-p position ?x)
     (has-feature-p position ?y)
     (decreasing-p (distance ?x ?y)))
```

where `has-feature-p` and `decreasing-p` are predefined predicates and `distance` could be a newly evolved detector. The symbols starting with a question mark represent variables, to be bound to items for activation of the template: the template can be activated when it can find a binding for its activation slot consisting of items that, when filled in, make the activation predicate become true. The action slot of a template could for example be to create a new item.

Features also have a history that reflects the change the feature value has made before it got its current value and is the way our event detection system handles time and change. Because it is impossible to record every change, these should be filtered and abstracted. We therefore adopted some ideas from qualitative physics (see e.g. [21, 10]): only the direction of change of a feature's value is recorded.

4 Semantic descriptions

The conceptualization module recognizes and filters events from the huge stream of raw data. In turn, the resulting event stream is filtered for events that are important to the agent at a certain time. This is the semantic subsystem's responsibility: to act as an attention focusing system, providing a concrete representation (semantic descriptions) for those aspects of the perceived environment that require immediate processing (action planning, verbalization, etc.)

4.1 Semantic descriptions

In our system, semantic descriptions (SD) express a certain aspect of the environment. The SD's are a form of predicate logic. Variables can correspond to any *item* from the conceptualization module, and thus to both objects and events. Predicates describe properties of an item, or relations betweens items. For reasons that will become clear in section 5 one variable is singled out as the head of a SD. For example "*the blue square approaches the circle.*" is described by the SD:

```
(?x | (approach ?x) (agent ?y ?x) (blue ?y)
      (square ?y) (patient ?z ?x) (circle ?z))
```

This sequence says that `?x` must be an approach event with an agent `?y` that is blue and square and a patient `?z` which is a circle. Variable `?x` is the head of the meaning; this is indicated by the explicit mention to the left of the operations. The same predicate sequence can also be used to express the head `?y`, in which case a natural language rendering could be *"the blue square approached by the circle"*.

The predicates in a SD can also play a functional role, for example `(patient ?z ?x)` can, given a suitable binding for `?z`, extract the patient of `?z` and bind it to `?x`. The agent has an evaluator that is able to process these descriptions within the set of current perceptions, and construct a set of bindings for the variables in a description that make the description true.

4.2 Meaning Construction

Since our agent lives in a complex, changing world, we want it to be adaptive. Instead of providing an agent with a predefined set of SD's, we must include a mechanism that allows the agent to create new SD's on-the-fly, as it needs them to describe something. Whenever the agent has constructed a new description, it can give it a name and incorporate it in its repertoire of operations. The constructed description can thus become itself a possible building block for future descriptions.

There is an important interplay between the conceptualization and the construction of SD's. When semantic descriptions are used often, this indicates that they are important, which might trigger a process to move the detection of the meaning one stage earlier, to the conceptualization phase. This could be compared to e.g. learning to dance: in the beginning one has to consciously think of every step one makes, but after a while the whole process becomes fluent and automatic.

5 Grammar

The final component of the system, the grammar, deals with transforming a SD to natural language and vice versa. Many grammar formalisms exist, but for our purposes we needed one that can deal with our SD's and that supports the fundamental properties of language: compositionality, phrase structure and recursion. Categorial grammar is such a formalism. In a categorial grammar every element of a language (i.e. a word or a phrase) has a syntactic category assigned to it. In the simplest case, there are only two basic categories: `n` and `s` (for noun and sentence, respectively). All other categories can be constructed by combining the basic categories according to certain constraints.

For example, "block" is of the basic category `n`. Now, if we want to say "the block", we need "the" to be of a category that can be combined *to the right* (the constraint) with an `n` to result in something of `n`. In our notation, "the" is therefore of category `(n n r)`: it results in a category `n` (the first one), if it is followed by (constraint `r`) a phrase of category `n` (the second one). Similarly, we can define a verb "approaches" as something that produces a complete sentence `s` if it is both preceded (constraint `l`) and followed by something of category `n`: `((s n l) n r)`. In this pure form, categorial grammar is equivalent to context-free grammars, but it has the advantages that all grammatical rules remain implicit in the lexicon and that the `r` and `l` constraints can easily be extended with less rigid constraints needed for free word-order languages.

In the categorial grammar tradition the usual way to deal with the meanings of *combinations* of lexical entries, uses Church's lambda calculus [see e.g. 5, for a discussion]. To do so we have to extend the semantic description to include lambda (λ) terms. Lambda terms can be seen as listing the variables that still need to be substituted; they disappear when a complete semantic description is reached. The semantic description for "x approaches y" is (x and y still need to be filled in):

(1) : `(λ?x λ?y | ?x | (approach ?z) (arg1 ?z ?x) (arg2 ?z ?y))`

When applied to the SD (2) : `(?p | (circle ?p))`, the resulting description is

(3): `(λ?y | ?p | (approach ?z) (arg1 ?z ?p) (arg2 ?z ?y)) (circle ?p))`

I.e. the variable `?x` in (1) is replaced by the head of (2), and the `λ?x` is removed.

We have implemented a production algorithm and an interpretation algorithm that, given the proper lexicon, map a SD on a natural language expression and vice versa. These are fairly straightforward search algorithms:

production starts with a target SD; the system selects all partial matches in the lexicon and searches for a way to combine these entries that yields a correct sentence, with a semantics that is identical to the target description.

interpretation starts with a natural language sentence; the system finds all partial matches in the lexicon and searches a way to combine these entries such that it matches the complete sentence and yields a consistent interpretation (a semantic description, without λ's and with all variables bound).

6 An example of the system at work

In this section we give an example of how the system behaves on input from a simple simulation shown in fig. 2, where two red squares are moving in opposite direction. In this example, the agent could be the speaker in a language game. It therefore has to pick a subject from the simulation to talk about, find a semantic description for it and finally verbalize this description in a grammatical correct utterance.

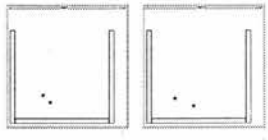


Figure 2. Two views of an example simulation.

The first step the system takes when a simulation starts is try to detect events. For this example, the system had definitions for various moving events (e.g. moving left, falling down), several kinds of objects (square, rectangle), various features (color) and some other events (approach). At the end of the frame sequence (see fig. 2), observation resulted in 5 body-items, 2 contact items, a move-left and move-right item, 2 falling items, 3 approach items and 3 move-away items.

The next step is to pick an item as a subject to talk about and find a (preferably unique) description for it. Suppose the system picks the square moving to the right. Some descriptions found by the system for this item were:

```
(?x | ((SQUARE ?x) (LOW ?x)))
(?x | ((MOVING-RIGHT ?y) (ARG1 ?y ?x)))
(?x | ((ARG2 ?y ?x) (MOVE-AWAY ?y) (ARG1 ?y ?u)
       (MOVING-LEFT ?u) (SQUARE ?x) (SQUARE ?y)))
```

The final step is to transform the semantic descriptions to a grammatical sentence. For this the system has a lexicon; the entry for the word "approaches" is as follows:

form	*λ's*	*head*	*meaning*	*category*
"approaches"	(λ?y λ?x)	`?x`	`((APPROACH ?z)` `(ARG1 ?z ?x) (ARG2 ?z ?y))`	((S N R) N L)

Using lexicon entries like this, the system is able to produce and interpret sentences or noun phrases that describe objects or events in a running simulation. The semantic descriptions given above would be translated to:

```
"the low square"
"the moving to the right"
"the red square moving away from the square moving to the left"
```

7 Discussion and future work

Our system is an attempt to combine ideas from both classical AI and new, adaptive AI. Both approaches have their difficulties and merits, both answer different questions. Often, the problems of one approach are the answers provided by the other. A major problem of classical AI is that it produces carefully engineered non-adaptive systems. A major problem of distributed, dynamically complex systems is that they are hard to engineer, unpredictable, and if a successful system is built it is hard to define what precisely it is that made it work.

An important thing that should be added to Harnad's list of challenges for cognitive theory [7] is how an agent can learn and adapt to changes in the environment. We argue language plays an important role in this and have designed a system that provides us with the means to test this hypothesis. At the same time, it allows us to investigate important issues on the origins and evolution of language.

While building an integrated, open system we had to make each subsystem powerful enough to meet other subsystems' requirements. For example, the semantic description language we developed needs to be able to handle all things it gets from the conceptualization module. It must also be able to provide the input required by the grammar module: if we want to incorporate tense and aspect in the language, the conceptualization module needs to know about time, in such a way that appropriate information can be propagated to the grammar module and vice versa. The principle of integration thus provided us at each level with some design guidelines.

But this principle can also be used by the system itself. If an agent feels at the grammatical or semantical level that a concept for an `approach+touch` event would be useful, it can instruct its conceptualization module to create a `collision` notion according to the requirements of the higher levels. We plan to use the system to investigate some specific aspects of language like tense, grammar, causality, etc. The main goal is to answer questions about the origins and evolution of language. We will therefore have to extend the system with good learning algorithms. In addition, we plan to replace the simulation by a camera and a robot arm. The simulation could still be used as an "imagination module" by an agent (see [17] for an implementation of this idea and related papers for psychological evidence for such a module in humans).

Acknowledgments

The authors would like to thank Luc Steels, who has created the theoretical framework underlying this research, and Frederic Vannieuwenhuyse and Eefje Leydesdorff for their contributions in the design and implementation of the grammar model. JDB is funded by the Vrije Universiteit Brussel (VUB), WZ is funded through the Concerted Research Action fund (G.O.A.) of the Flemish Government and the VUB. JVL is sponsored by a grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

[1] David Baraff. Fast contact force computation for nonpenetrating rigid bodies. *Computer Graphics*, 28(Annual Conference Series):23–34, 1994.

[2] J. Batali. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe, editor, *Linguistic evolution through language acquisition: formal and computational models.* Cambridge University Press, Cambridge, UK, 2002.

[3] A. Chatterjee and A. Ruina. A new algebraic rigid body collision law based on impulse space considerations. *ASME Journal of Applied Mechanics.*, 1998.

[4] Bart de Boer. *The origins of vowel systems.* Oxford University Press, Oxford, UK, 2001.

[5] L.T.F. Gamut. *Logic, language and meaning*, volume 2. University of Chicago Press, 1991.

[6] James K. Hanh. Realistic animation of rigid bodies. *Computer Graphics (SIGGRAPH '88 Proceedings)*, 22(4):299–308, 1988.

[7] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[8] James R. Hurford. Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, 77:189–222, 1989.

[9] Simon Kirby. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguistic evolution through language acquisition: formal and computational models.* Cambridge University Press, 2002.

[10] Benjamin Kuipers. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge.* MIT Press, Cambridge, MA, 1994.

[11] Bruce J. MacLennan and Gordon M. Burghardt. Synthetic ethology and the evolution of cooperative communication. *Adaptive Behavior*, 2(2):161–188, 1993.

[12] B. Mirtich. Hybrid simulation: Combining constraints and impulses. In *Proceedings of the First Workshop on Simulation and Interaction in Virtual Environments.*, 1995.

[13] Brian Mirtich and John F. Canny. Impulse-based simulation of rigid bodies. In *Symposium on Interactive 3D Graphics*, pages 181–188, 217, 1995.

[14] Matthew Moore and Jane Wilhelms. Collision detection and response for computer animation. *Computer Graphics*, 22(4), 1988.

[15] J. Noble and D. Cliff. On simulating the evolution of communication. In P. Maes et al, editor, *SAB96*, Cambridge MA, 1996. MIT Press.

[16] M. Oliphant. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 7(3/4), 1999.

[17] Jeffrey Mark Siskind. Naive physics, event perception, lexical semantics, and language acquisition. Ph.D. diss. (Elec. Eng'g. & Comp. Sci., MIT)., 1992.

[18] L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. Crucial factors in the origins of word-meaning. In A. Wray, editor, *The Transition to Language.* Oxford University Press, Oxford, UK, 2002.

[19] Luc Steels. Emergent adaptive lexicons. In P. Maes, editor, *Proceedings of the Simulation of Adaptive Behaviour Conference.* The MIT Press, Cambridge, Ma., 1996.

[20] Luc Steels. The origins of syntax in visually grounded robotic agents. In M. Pollack, editor, *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI97) (Los Angeles, California)*, San Francisco, CA, 1997. Morgan Kauffman Publishers.

[21] Daniel S. Weld and Johan de Kleer. *Readings in Qualitative Reasoning About Physical Systems.* Morgan Kaufmann, San Mateo, CA, 1990.

[22] Terry Winograd. *Understanding Natural Language.* Academic Press, 1976.

Language adaptation helps language acquisition

Willem Zuidema
AI Lab – Vrije Universiteit Brussel
Pleinlaan 2, B-1050, BRUSSELS, Belgium
jelle@arti.vub.ac.be

Abstract

Language acquisition is a very particular type of learning problem: it is a problem where the target of the learning process is itself the outcome of a learning process. Language can therefore adapt to the learning algorithm. I present a model that shows that due to this effect – and contrary to some claims from the Universal Grammar tradition – "unlearnable" grammars can be successfully acquired, and grammatical coherence in a population can be maintained.

1 Introduction

Human language is one of the most intriguing adaptive behaviors that has emerged in evolution. Language makes it possible to express an unbounded number of different messages, and it serves as the vehicle for transmitting knowledge that is acquired over many generations. Not surprisingly, the origins of language are a central issue in both evolutionary biology and the cognitive sciences.

The dominant explanation for the origins and nature of human language postulates a "Universal Grammar": an innate system of principles and parameters, that is universal, genetically specified and independent from other cognitive abilities. In this paper, I study an argument that lies at the heart of this dominant position: the argument from the poverty of stimulus. This argument states that children have insufficient evidence to learn the language of their parents without innate knowledge about which languages are possible and which are not. This claim is backed-up with a series of mathematical models. Here, we will focus our discussion on two such models: Gold (1967) and Nowak et al. (2001).

Gold (1967) introduced the criterion "identification in the limit" for evaluating the success of a learning algorithm: with an infinite number of training samples all hypotheses of the algorithm should be identical, and equivalent to the target. Gold showed that context-free grammars are in general not learnable by this criterion from positive samples alone. This proof is based on the fact that if one has a grammar G that is consistent with all the training data, one can always construct a grammar G' that is slightly more general: i.e. the language of G, $L(G)$ is a subset of $L(G')$.

Nowak et al. (2001) provide a novel variant of the argument from the poverty of stimulus, that is based on a mathematical model of the evolution of grammars. The first step of their argument is a "coherence threshold". This threshold is the minimum learning accuracy of an individual that is consistent with grammatical coherence in a population, i.e. with a majority of individuals to use the same grammar. The second step relates this coherence threshold to a lower bound (b_0) on the number of sample sentences that a child needs. They derive that b_0 is proportional to the total number of possible grammars N. From this and the fact that the number of sample sentences is finite, Nowak et al. conclude that only if N is relatively small can a stable grammar emerge in a population. I.e. the population dynamics require a restrictive Universal Grammar.

2 Model design

These models have in common that they implicitly assume that every possible grammar is equally likely to become the target grammar for learning. If even the best possible learning algorithm cannot learn such a grammar, the set of allowed grammars must be restricted. There is, however, reason to believe that this assumption is not the most useful for language learning. Language learning is a very particular type of learning problem, because the outcome of the learning process at one generation is the input for the next.

The model study I present here is motivated by this observation. The model consists of an evolving population of language learners, that learn a grammar from their parents and get offspring proportional to the success in communicating with other individuals in their generation. The grammar induction procedure is fixed; it is inspired by Kirby (2000). The details of the grammatical formalism (context-free grammars) and the population structure are deliberately close to Gold (1967) and Nowak et al. (2001) respectively.

I use context-free grammars to represent the linguistic abilities. In particular, the representation is limited to grammars G where all rules are of one of the following forms: $A \mapsto t$, $A \mapsto BC$, or $A \mapsto Bt$. Since

every context-free grammar can be transformed to such a grammar, the restrictions on the rule-types above do not limit the scope of languages that can be represented. They are, however, relevant for the language acquisition algorithm that will be discussed below. Note that the class of languages that the formalism can represent is unlearnable by Gold's criterion.

The language acquisition algorithm used in the model consists of three operations: (i) incorporation (extend the language, such that it includes the encountered string), (ii) compression (substitute frequent and long substrings with a nonterminal, such that the grammar becomes smaller and the language remains unchanged), (iii) generalization (equate two nonterminals, such that the grammar becomes smaller and the language larger).

3 Results

The main result is in figure 1, which shows two curves: (i) the average communicative success of agents speaking with their parents which is the measure for the *learnability* of the language (labeled "between generation C"), and (ii) the average communicative success of agents speaking with other agents of the same generation (labeled "within generation C") which gives the fitness of agents and is a measure for the grammatical variation in the population.

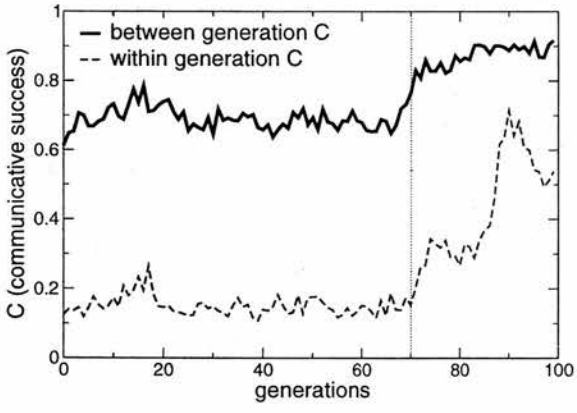


Figure 1: *Parameters are:* $V_t = \{0,1,2,3\}$, $V_{nt} = \{S,a,b,c,d,e,f\}$, *P=20, T=100, M=100,* l_0*=12*

For a long period the learning is not very successful. The between generation C is low (grammars are unlearnable), and consequently the within generation C is also low (the dynamics are below the "coherence threshold" of Nowak et al. 2001). In other words, individuals are so bad at learning that members of the population can not understand each other. Around generation 70 this situation suddenly changes. The between generation C rises, and very quickly also the within generation C rises to non-trivial levels. With always the same number of sample sentences, and with always the same grammar space, there are regions of that space where the dynamics are apparently under the coherence threshold, while there are other regions where the dynamics are above this threshold. The language has adapted to the learning algorithm, and, consequently, the coherence does not satisfy the prediction of Nowak et al. In many runs (not shown here) I have also observed 100% learning accuracy of children. The grammars in this situations are thus learnable by Gold's criterion. In some, but not all cases, these emergent grammars are recursive.

4 Discussion

I believe that these results, simple and preliminary as they may be, have some important consequences for our thinking about language acquisition. In studies like the mathematical models of Gold and Nowak et al., one derives from the properties of the learning procedure (the search procedure), fundamental constraints on the nature of the target grammar (the search space). My results, like those of Kirby (2000) and others, indicate that in *iterated learning* it is not necessary to put the (whole) explanatory burden on constraints on the search space. In my model, the target grammars are learnable, not because the used formalism imposes restrictions on the grammars, but because the targets dynamically change and – in the iteration of learners learning from learners – adapt to the used learning algorithm. In other words, neither the search space nor the search procedure directly determine which grammars "exist"; the set of target grammars at the end of the simulation is the emergent result of iterating a search process over and over again.

Isn't this Universal Grammar in disguise? Learnability is – consistent with the undisputed proof of Gold (1967) – still achieved by constraining the set of targets. However, unlike in usual *interpretations* of this proof, these constraints are not strict (some grammars are better learnable than others, allowing for an infinite "Grammar Universe"), and they are not a-priori: they are the outcome of iterated learning. The poverty of stimulus is here no longer a problem; instead, the ancestors' poverty is the solution for the child's.

References

Gold, E. M. (1967). Language identification in the limit. *Information and Control (now Information and Computation)*, 10:447–474.

Kirby, S. (2000). Syntax without natural selection. In Knight, C., Hurford, J., and Studdert-Kennedy, M., (Eds.), *The Evolutionary Emergence of Language.* Cambridge University Press.

Nowak, M. A., Komarova, N., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291:114–118.

Optimal communication in a noisy and heterogeneous environment

Willem Zuidema

Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute for Cell, Animal and Population Biology
University of Edinburgh
40 George Square, Edinburgh EH8 9LL
Scotland, United Kingdom
jelle@ling.ed.ac.uk
http://www.ling.ed.ac.uk/~jelle

Abstract. Compositionality is a fundamental property of natural language. Explaining its evolution remains a challenging problem because existing explanations require a structured language to be present before compositionality can spread in the population. In this paper, I study whether a communication system can evolve that shows the preservation of topology between meaning-space and signal-space, without assuming that individuals have any prior processing mechanism for compositionality. I present a formalism to describe a communication system where there is noise in signaling and variation in the values of meanings. In contrast to previous models, both the noise and values depend on the topology of the signal- and meaning spaces. I consider a population of agents that each try to optimize their communicative success. The results show that the preservation of topology follows naturally from the assumptions on noise, values and individual-based optimization.

1 Major transitions in the evolution of language

Human languages are unique communication systems in nature because of their enormous expressiveness and flexibility. They accomplish this by using combinatorial principles in phonology, morphology and syntax [8], which impose important requirements on the cognitive abilities of language users. Explaining the origins of the structure of language and the human abilities to process it is a challenging problem for linguistics, cognitive science and evolutionary biology. Mathematical and computational models have been invaluable tools for getting a grip on this problem [11].

Jackendoff [8] has laid out a scenario for the various stages in the evolution of human language from primate-like communication, that reflects a growing consensus and can be summarized with the following "major transitions":

1. From situation-specific signals (e.g. alarm calls), to signals that are non-situation-specific but from a closed class;

2. From (1) to an open, unlimited (learned) class of signals and, subsequently, a phonological combinatorial system;
3. From (1) to the concatenation of signals and, subsequently, the use of ordering of signals to convey semantic relations ("compositionality");
4. From (2) and (3), which constitute the ingredients of a protolanguage, to hierarchical phrase structure and recursion,
5. From (4) to modern language, with a vocabulary for abstract semantic relations, grammatical categories, grammatical functions and a complex morphology.

Presumably, all transitions have greatly increased the number of distinct "signs" (signal–meaning pairs) that can be expressed, transmitted, memorized and learned. Jackendoff argues convincingly that modern languages contain "fossils" of each of the intermediate stages. E.g. the compound noun construction in English can be viewed as a fossil of stage (3): the meaning of words like "doghouse" and "housedog" is deducible (but not completely specified) from the meaning of the component words and the order in which they are put.

Less consensus exists on how the transition from each stage to another could have happened. Some have argued for extensive innate, language-specific cognitive specializations that have evolved under natural selection (e.g. [1, 8]. This is an appealing position, in line with dominant "nativist" theories in linguistics and evolutionary biology. Unfortunately, explanations of this type have generally remained much *underspecified.* Jackendoff admits: *"I will not inquire as to the details of how increased expressive power came to spread through a population [...]. Accepted practice in evolutionary psychology [...] generally finds it convenient to ignore these problems."([8], p. 237)*

Ignoring this problem is an unfortunate tradition. Understanding how innovations can spread in a population is the essence of any evolutionary explanation, and a better end-result is neither a sufficient nor a necessary condition for the spread of innovations. Specifically in the case of language, the spread of innovations is not at all obvious, even if the end-result – when the whole population has adopted an innovation – is demonstrably better, because of two important difficulties that arise from the *frequency-dependency* of language evolution : (i) if only the hearers benefit from communication, it is not clear why speakers would evolve as to give away – altruistically – more and more information [15, 21]; (ii) even if both speakers and hearers benefit, it is not clear how there can be a positive selection pressure on a linguistic innovation if that innovation appears in a population that uses a language without it, and, moreover, how that pressure can be strong enough to prevent it from being lost by drift [6, 3, 21].

A number of researchers have explored the possibilities of general learning and cognitive abilities and cultural evolution explaining the transitions instead (see [20, 11] for reviews and references), or, of cultural evolution facilitating the genetic evolution of linguistic innovations [5, 9]. These models are useful in clarifying the conditions for the "major transitions", but face some new difficulties themselves as well: (i) in many cases, the assumed cognitive abilities are much more language-specific than one would like; (ii) cultural evolution, such as

the progressively better structured languages in the "Iterated Learning Model" [10, 2], only takes off when there is already some initial, random structure in the language.

Explaining the evolution of aspects of natural language like combinatorial phonology and compositionality thus remain challenging problems because both the genetic and the cultural evolution explanation require a structured language to be already present in the population before the linguistic innovations can successfully spread in a population. In this paper, I focus on compositionality: the property that the meaning of the whole (e.g. a sentence) is a function of the meaning of the parts (e.g. the words) and the way they are put together. I do not study the evolution of compositionality itself, but explore a possible route for a structured language to emerge without the capacity for compositionality present in the population. That structure is *topology preservation* between meaning-space and signal-space, i.e. similar meanings are expressed with similar signals.

In the next section I present a formalism to describe a communication system where there is noise in signaling and variation in the values of meanings. In contrast to previous models, both the noise and values depend on the topology of the signal- and meaning spaces. In section 3 I present a model of a population of agents that each try to optimize their communicative success under these circumstances. The results, in section 4, show that the preservation of topology between meaning-space and signal-space follows naturally from the assumptions on noise, values and individual-based optimization.

2 A formalism for communication under noisy conditions

Assume that there are M different meanings that an individual might want to express, and F different signals (forms) that it can use for this task. The communication system of an individual is represented with a *production matrix* $\mathbf{S}$ and an *interpretation matrix* $\mathbf{R}$. $\mathbf{S}$ gives for every meaning m and every signal f, the probability that the individual chooses f to convey m. Conversely, $\mathbf{R}$ gives for every signal f and meaning m, the probability that f will be interpreted as m. $\mathbf{S}$ is thus a $M \times F$ matrix, and $\mathbf{R}$ a $F \times M$ matrix. Variants of this notation are used by [7, 14] and other researchers.

In addition, following [13], I assume that signals can be more or less similar to each other and that there is noise on the transmission of signals, which depends on these similarities. Further, I assume that meanings can be more or less similar to each other, and that the value of a certain *interpretation* depends on how close it is to the *intention*. These aspects are modeled with a *confusion matrix* $\mathbf{U}$ (of dimension $F \times F$) and a *value matrix* $\mathbf{V}$ (of dimension $M \times M$). This notation is an extension of the notation in [13], and was introduced in [22].

These four matrices together can describe the most important aspects of a communication system: which signals are used for which meanings by hearers and by speakers, how likely it is that signals get confused in the transmission, and what the consequences of a particular successful or unsuccessful interpretation are. Interestingly, they combine elegantly in one simple expression for the

expected payoff w_{ij} of communication between a hearer i and a speaker j [22]:

$$w_{ij} = \mathbf{V} \cdot (\mathbf{S}^i \times (\mathbf{U} \times \mathbf{R}^j)) \tag{1}$$

In this formula, "×" represents the usual matrix multiplication and "·" represents dot-multiplication (the sum of all multiplications of corresponding elements in both matrices; the result of dot-multiplication is not a matrix, but a scalar).

A hypothetical example, loosely based on the famous Vervet monkey alarm calls [17], might make the use of this formalism and measure clear. Imagine an alarm call system of a monkey species for three different types of predators: from the air (eagles), from the ground (leopards) and from the trees (snakes). Imagine further that the monkeys are capable of producing a number (say 5) of different sounds that range on one axis (e.g. pitch, from high to low) and tahte these are more easily confused if they are closer together. Thus, the confusion matrix $\mathbf{U}$ might look like in the left matrix of figure 1.

$U =$

	received signal				
sent signal ↓	1*kHz*	2*kHz*	3*kHz*	4*kHz*	5*kHz*
1*kHz*	0.7	0.2	0.1	0.0	0.0
2*kHz*	0.2	0.6	0.2	0.0	0.0
3*kHz*	0.0	0.2	0.6	0.2	0.0
4*kHz*	0.0	0.0	0.2	0.6	0.2
5*kHz*	0.0	0.0	0.1	0.2	0.7

$V =$

	intentions		
interpretations ↓	*eagle*	*snake*	*leopard*
eagle	0.9	0.5	0.1
snake	0.2	0.9	0.2
leopard	0.1	0.5	0.9

Fig. 1. Confusion and value matrices for the monkeys in the example, describing the noise in signaling and the value of intention–interpretation pairs in their environment.

Further, although it is obviously best to interpret a signal correctly, if one makes a mistake, typically not every mistake is equally bad. For example, if a leopard alarm is given, the leopard response (run into a tree) is best, but a snake response (search surrounding area) is better than an eagle response (run into a bush, where leopards typically hide) [17]. Thus the value matrix $\mathbf{V}$ might look something like the right matrix in figure 1.

$S =$

	sent signal				
intention ↓	1*kHz*	2*kHz*	3*kHz*	4*kHz*	5*kHz*
eagle	1.0	0.0	0.0	0.0	0.0
snake	0.0	0.0	1.0	0.0	0.0
leopard	0.0	0.0	0.0	0.0	1.0

$R =$

	interpretation		
received signal ↓	*eagle*	*snake*	*leopard*
1*kHz*	1.0	0.0	0.0
2*kHz*	1.0	0.0	0.0
3*kHz*	0.0	1.0	0.0
4*kHz*	0.0	0.0	1.0
5*kHz*	0.0	0.0	1.0

Fig. 2. Production and interpretation matrices for the monkeys in the example, describing which signals they use for which situations.

For any given production and interpretation matrix, we can through equation (1) calculate the expected payoff from communication. Assume a speaker i with its $\mathbf{S}^i$ as the left matrix in fig. 2, and a hearer j with its $\mathbf{R}^j$ as the right matrix in that figure. The expected payoff of the interaction between i and j if the constraints on communications are as in $\mathbf{U}$ and $\mathbf{V}$ in fig. 1 is, by proper application of equation (1): $w_{ij} = 0.7 \times 0.9 + 0.2 \times 0.5 + 0.2 \times 0.5 + 0.6 \times 0.9 + 0.2 \times 0.5 + 0.1 \times 0.5 + 0.2 \times 0.9 + 0.7 \times 0.9 = 2.33$

In this simple example, the matrices $\mathbf{U}$ and $\mathbf{V}$ are very small, and reflect only a 1-dimensional topology in both signal and meaning space. The matrices $\mathbf{S}$ and $\mathbf{R}$ are set by hand to arbitrarily chosen values. In contrast, in the simulations of this paper we will consider larger and more complex choices for $\mathbf{U}$ and $\mathbf{V}$, and we will use a hill-climbing algorithm to find the appropriate (near-) optimal settings for $\mathbf{S}$ and $\mathbf{R}$.

3 Distributed hill-climbing

Based on the measure of equation (1), I use a hill-climbing algorithm to improve the communication. To speed up the simulations, I make the simplification that the values in the $\mathbf{S}$ and $\mathbf{R}$ matrices are all either 1 or 0, i.e. they are deterministic encoders and decoders, which can be shown to always perform better than their stochastic versions [18, 16]. Hill-climbing in the simulations reported here is *distributed*, i.e. I simulate a population (size 400) of agents that each try to optimize their success in communicating with a randomly selected other agent (see the author's website for details). Experiments (not reported here) suggest that distributed hill-climbing, although orders of magnitude faster, leads to very similar results as global hill-climbing. In some of the simulations, agents are placed on a grid (size 20×20) and interact only with their direct neighbors (8, except for agents at the edge which have less neighbors), but also in this condition very similar results are obtained.

The motivation for this style of optimization is (i) that it is fast and straightforward to implement; (ii) that it works well, and gives, if not the optimum, a good insight on characteristics of the optimal communication system; and (iii) that it shows possible *routes* to (near-) optimal communication systems, and in a sense forms an abstraction for both learning and evolution.

The $\mathbf{V}$ and $\mathbf{U}$ matrices can be chosen to reflect all kinds of assumptions about the signal and meaning space. In this paper I vary whether all meanings are equally valuable ($v = 1.0$, labeled "homogeneous"), or get assigned a random value ($0.0 < v \leq 1.0$, labeled "heterogeneous"). I further vary whether or not there is a topology, and if so, of which dimensionality. The diagonal elements of $\mathbf{V}$ are always v and of $\mathbf{U}$ always 1.0. Without a topology ("0d"), the off-diagonal elements in $\mathbf{U}$ or $\mathbf{V}$ are 0. With a topology, the off-diagonal elements are given by $\mathbf{V}(p,q) = v/(1 + D(p,q))$ and $\mathbf{U}(p,q) = 1/(1 + D(p,q))$, where $D(p,q)$ gives the squared Euclidean distance between the positions of the two meanings or signals i and j. In the 1-dimensional condition ("1d"), the position of a meaning or signal is simply defined as its index. In the 2d condition, the meaning and signal spaces are 2-dimensional surfaces of size $(\sqrt{M} \times \sqrt{M})$ or $(\sqrt{F} \times \sqrt{F})$. The x-coordinate is then given by the largest integer smaller than the root of the index: $x = int(\sqrt{i})$. The y-coordinate by: $y = i \ modulo \ x$. After these values are set, the rows of both $\mathbf{U}$ and $\mathbf{V}$ matrices are normalized.

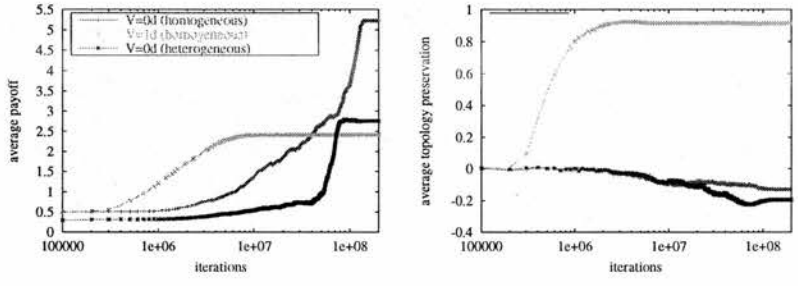
I monitor the behavior of the model with two measures. The first is the average payoff, as given by equation (1), averaged over all individuals interacting with all other individuals, both as speaker and as hearer. The second is a measure

for the degree of topology preservation between the meaning space and the signal space in the emerging languages. Following [2], I use the correlation ("Pearson's r") between the distance between each pair of meanings and the distance between the corresponding signals:

$$r = \underset{m,m' \in M}{correlation} \left(D\left(m, m'\right), D\left(S[m], S[m']\right) \right), \tag{2}$$

where $S[m]$ gives the most likely signal used to express m according to $\mathbf{S}$.

For 2-dimensional meaning spaces I also visualize the topology preservation by plotting all meanings as nodes in a meaning space, and connecting those nodes where the corresponding signals are (one of maximal 4) neighbors in signal-space.



(a) Payoff ($\overline{w}$) (b) Topology Preservation ($\overline{r}$)

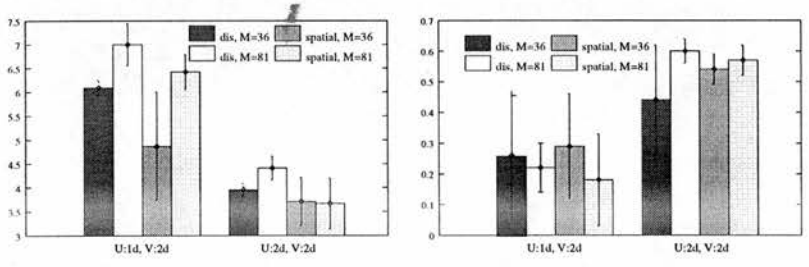
Fig. 3. Average payoff (a) and degree of topology preservation (b) for 2×10^8 iterations under 3 conditions: (1) **V**:0d homogeneous, (2) **V**:1d homogeneous; (3) **V**:0d heterogeneous. The maximum average payoffs that are reached depend on the arbitrary chosen values of the **V** matrices; hence, only the shapes of the curves are important. Common parameters are P=400, M=16, F=49, **U**:1d.

4 Results

Figure 3 shows the average payoff and topology preservation for simulations under 3 different conditions: (i) homogeneous and no topology in the meaning space ("**V**:0d"); (ii) homogeneous and **V**:1d; (iii) heterogeneous and **V**:0d. The results are plotted with a logarithmic x-axis. They show that convergence is more than 10 times faster if there is a topology in the meaning space. Recall that in the topology condition, interpretations with a meaning close to the intention are also rewarded. That fact facilitates establishing conventions regarding which signals to use for which meanings, because it creates more possibilities to break the initial symmetry (when no convention is established, every signal–meaning pair is equally good or bad).

Figure 4 shows the average payoff and topology preservation for 60 simulations where the dimensionality of the signal space is varied, and where hearers

are selected randomly from either the whole population ("dis"), or from one of the speaker's 8 neighbors ("spatial"). In all cases, the payoff reaches high levels (when the signal space is 1d) or intermediate levels (when the signal space is 2d and the overall noise-level is consequently higher because each signal has more neighbors). Also, in all cases the topology preservation reaches high levels (when the dimensionalities of meaning and signal space match) or intermediate levels (when the dimensionalities mismatch).



(a) Payoff ($\overline{w}$) (b) Topology preservation ($\overline{r}$)

Fig. 4. Average payoff (a) and degree of topology preservation (b) after 5×10^7 iterations for different parameters. Error-bars indicate standard deviations. Common parameters are P=400, M=36 and **V**:2d heterogeneous.

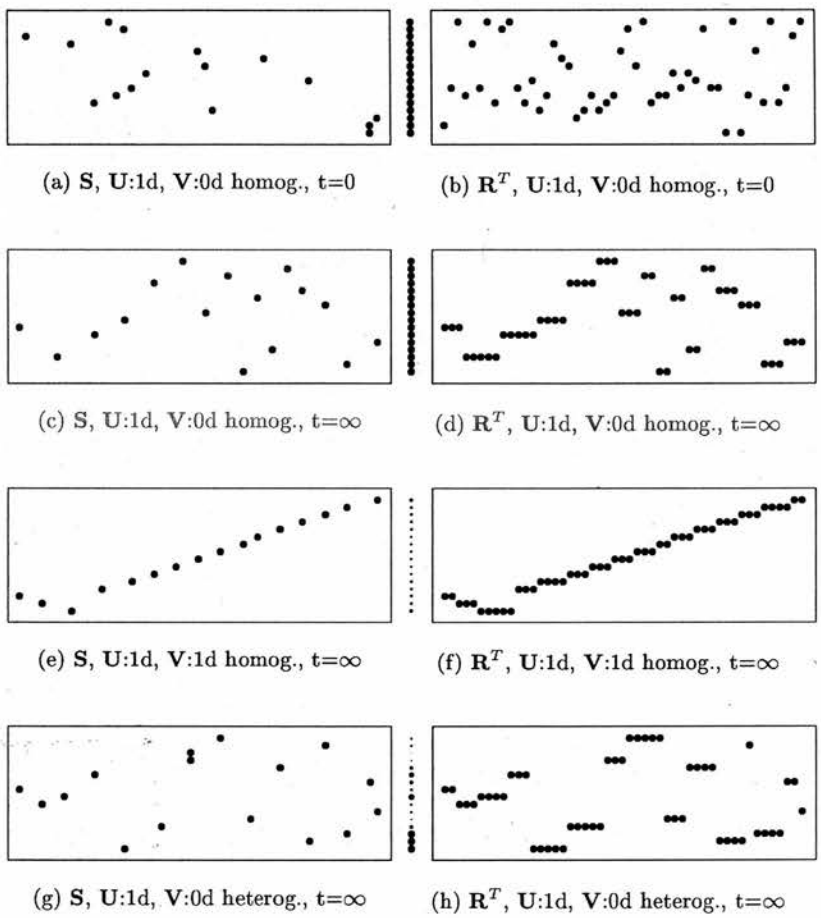
The emerging communication systems are visualized in fig. 5 and 6 and can be summarized with the following properties:

Specificity: every meaning has exactly one signal to express it and vice versa (i.e. no homonyms, and no real synonyms: if different signals have the same meaning they are very similar to each other).

Coherence: all agents agree on which signals to use for which meanings, and vice versa. Specificity and coherence are also found in "language game" models where there is no noise on signaling (e.g. [14, 19]).

Distinctiveness: in the **S** matrices, the used signals are maximally dissimilar to each other, so that they can be easily distinguished (compare figure 5a, at the start of the simulation, with 5c, at equilibrium). In the **R** matrices, clusters of neighboring signals all are interpreted as the same meaning. Typically, the most central signal (except at the edges) in such a cluster is the one that is actually used by the **S** matrix (compare figure 5c with 5d). Distinctiveness is also found in the "imitation game" [4], where no meanings are modeled.

Topology preservation: if there is a topology in both the meaning- and signal-space (as determined by **V** and **U**), similar signals tend to have similar meanings [22]. This preservation is not perfect (there is one major irregularity and several minor ones in the signal–meaning mapping of figure 5e and f. The topology preservation, according to equation (2), is $\overline{r} = 0.915$), but in all simulations performed it is surprisingly high. "Bad" solutions, such as the **S** and **R** of figures 5c



(a) **S**, **U**:1d, **V**:0d homog., t=0

(b) $\mathbf{R}^T$, **U**:1d, **V**:0d homog., t=0

(c) **S**, **U**:1d, **V**:0d homog., t=∞

(d) $\mathbf{R}^T$, **U**:1d, **V**:0d homog., t=∞

(e) **S**, **U**:1d, **V**:1d homog., t=∞

(f) $\mathbf{R}^T$, **U**:1d, **V**:1d homog., t=∞

(g) **S**, **U**:1d, **V**:0d heterog., t=∞

(h) $\mathbf{R}^T$, **U**:1d, **V**:0d heterog., t=∞

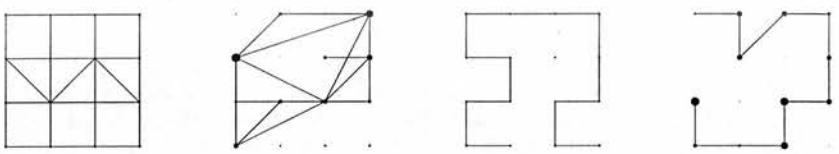
Fig. 5. (a)-(h) Examples of **S** and **R** matrices from the simulations of figure 3. For easy comparison, the **R** matrices are transposed so that in all matrices meanings differ on the vertical axis, and signals on the horizontal axis. Between the matrices the diagonal values of the **V** matrix are plotted, where the diameter of a circle corresponds to value of the corresponding meaning. Common parameters are P=400, M=16, F=49.

and d ($\overline{r} = -0.073$), are stable once established in the population, but have a much smaller basin of attraction. In the case of a two-dimensional meaning space, we can draw plots like figures 6a-d, which show that the topology is almost perfectly preserved if the dimensionalities of the meaning- and signal-spaces match (6a), although it is skewed if different meanings receive very different values

(6b). But even if the dimensionalities do not match, there is a strong tendency to preserve topology as well as possible (6c and d).

Valuable meanings first: When one analyzes the intermediate stages between the random initialization and the equilibrium solutions (not shown here; see author's website), it becomes clear that with a heterogeneous **V** valuable meaning-signal pairs get established first, and change little afterward.

Meanings sacrificed: Finally, when the **V** matrix is heterogeneous (figure 6b and d), or there is a dimensionality-mismatch (figure 6c and d), one can observe that meanings with very low value are sacrificed for the benefit of robust recognition of more valuable meanings (a similar observation was made in [13]). These sacrificed meanings "deliberately" get expressed with a signal that will be interpreted with a meaning that is very close.



(a) **U**:2d, **V**:2d ho., $\overline{r} = 0.67$

(b) **U**:2d, **V**:2d he., $\overline{r} = 0.65$

(c) **U**:1d, **V**:2d ho., $\overline{r} = 0.46$

(d) **U**:1d, **V**:2d he., $\overline{r} = 0.07$

Fig. 6. Topology preservation at equilibrium in 4 simulations with 1d and 2d **U** matrices, and homogeneous and heterogeneous 2d **V** matrices. Nodes are meanings (diameters correspond to value), edges connect neighbors in signal space (several signals can map to a single meaning, such that nodes can have many neighbors; some meanings are not expressed, and the corresponding nodes are not connected). Common parameters are P=400, M=16, F=49.

5 Conclusions

In this paper, I have shown that from simple assumptions about topologies in meaning- and signal-space, and individual-based optimization, communication systems can arise that show a structured mapping from meanings to signals. In a population where such a language is spoken, the fundamental new phenomenon of compositionality can presumably much more easily evolve.

There is no space here to explore the many connection between these simulations and the fields of Information Theory [16] and Evolutionary Game Theory [12]. In a sense, the matrices of figure 5 and 6 describe evolutionary stable strategies, under the constraints of communication over a noisy channel. These connections, and the analytic proofs that can be worked out in these frameworks, will be the topic of future work.

Acknowledgments Funding from the Prins Bernhard Cultuurfonds and a Marie Curie fellowship of the European Commission is gratefully acknowledged. I thank Gert Westermann and Andy Gardner for their remarks.

References

[1] D. Bickerton. *Language and Species.* University of Chicago Press, 1990.
[2] H. Brighton. *Simplicity as a Driving Force in Linguistic Evolution.* PhD-thesis, Theoretical and Applied Linguistics, University of Edinburgh, 2003.
[3] L.L. Cavalli-Sforza and M.W. Feldman. Paradox of the evolution of communication and of social interactivity. *Proc. Nat. Acad. Sci. USA*, 80:2017–2021, 1983.
[4] B. de Boer. Self organization in vowel systems. *J. of Phonetics*, 28:441–465, 2000.
[5] T. Deacon. *Symbolic species, the co-evolution of language and the human brain.* The Penguin Press, 1997.
[6] R.A. Fisher. On the dominance ratio. *Proc Roy Soc Edin*, 42:321–431, 1922.
[7] J. Hurford. Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222, 1989.
[8] R. Jackendoff. *Foundations of Language.* Oxford University Press, 2002.
[9] S. Kirby and J. Hurford. Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands and I. Harvey, editors, *Proceedings 4th European Conference on Artificial Life*, pages 493–502. MIT Press, Cambridge, MA, 1997.
[10] S. Kirby. Syntax without natural selection. In C. Knight et al., editors, *The Evolutionary Emergence of Language.* Cambridge University Press, 2000.
[11] S. Kirby. Natural Language from Artificial Life. *Artificial Life*, 8(2):185–215, 2002.
[12] J. Maynard Smith. *Evolution and the Theory of Games.* Cambridge University Press, Cambridge, UK, 1982.
[13] M.A. Nowak and D.C. Krakauer. The evolution of language. *Proc. Nat. Acad. Sci. USA*, 96:8028–8033, 1999.
[14] M. Oliphant and J. Batali. Learning and the emergence of coordinated communication. *Center for research on language newsletter*, 11(1), 1996.
[15] M. Oliphant. The dilemma of Saussurean communication. *BioSystems*, 37(1-2):31–38, 1994.
[16] J.B. Plotkin and M.A. Nowak. Language evolution and information theory. *Journal of Theoretical Biology*, pages 147–159, 2000.
[17] R.M. Seyfarth and D.L. Cheney. Some general features of vocal development in nonhuman primates. In C.T. Snowdon and M. Hausberger, editors, *Social influences on vocal development*, pages 249–273. Cambridge University Press, 1997.
[18] C.E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423 and 623–656, 1948.
[19] L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. Crucial factors in the origins of word-meaning. In A. Wray, editor, *The Transition to Language.* Oxford University Press, Oxford, UK, 2002.
[20] L. Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1:1–35, 1997.
[21] W. Zuidema and P. Hogeweg. Selective advantages of syntactic language: a model study. In Gleitman and Joshi, editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 577–582. Lawrence Erlbaum, 2000.
[22] W. Zuidema and G. Westermann. Evolution of an Optimal Lexicon under Constraints from Embodiment. *Artificial Life*, 2003. (accepted).

Evolution of an Optimal Lexicon under Constraints from Embodiment

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
Institute for Cell, Animal and Population Biology
University of Edinburgh
40 George Square
Edinburgh EH8 9LL
United Kingdom
jelle@ling.ed.ac.uk

Gert Westermann
Centre for Brain and Cognitive Development
Birkbeck College
University of London
Malet Street
London WC1E 7HX
United Kingdom
g.westermann@bbk.ac.uk

Abstract Research in language evolution is concerned with the question of how complex linguistic structures can emerge from the interactions between many communicating individuals. Thus it complements psycholinguistics, which investigates the processes involved in individual adult language processing, and child language development studies, which investigate how children learn a given (fixed) language. We focus on the framework of *language games* and argue that they offer a fresh and formal perspective on many current debates in cognitive science, including those on the synchronic-versus-diachronic perspective on language, the embodiment and situatedness of language and cognition, and the self-organization of linguistic patterns. We present a measure for the quality of a lexicon in a population, and derive four characteristics of the optimal lexicon: specificity, coherence, distinctiveness, and regularity. We present a model of lexical dynamics that shows the spontaneous emergence of these characteristics in a distributed population of individuals that incorporate embodiment constraints. Finally, we discuss how research in cognitive science could contribute to improving existing language game models.

Keywords
Language, lexicon, evolution, self-organization, embodiment

1 Introduction

There exists a long tradition of formulating and studying formal models of language processing and language learning. These models have generally focused on the linguistic competence of a single individual. They have proven to be appealing because such formalisms offer precision and clarity, have led to successful technology, and have allowed for extensive theoretical research to complement empirical work.

However, these competence models have abstracted away many arguably crucial characteristics of language. These abstractions are viewed with growing uneasiness by cognitive scientists, linguists, and other researchers. Some of their concerns are well known: competence theories lack an appreciation of linguistic performance and of the communicative function of language, and they place a strong emphasis on symbolic processing and innateness (see, e.g., [8, 33, 17] for criticisms).

Here we focus on a particular criticism: traditional models fail to acknowledge how much of linguistic structure emerges from communication and embodiment. Recent research on natural language pragmatics, for instance, has focused on language as a cooperative phenomenon where communication is viewed as a *joint action* between the participants [4]. This view is in contrast to the traditional approach in which speaking and hearing are investigated in isolation as *individual actions*. Researchers in the

framework of *emergentism* have argued that the structure of language should be explained as the emergent result of the many interactions between known processes in evolution, development, speaking, listening, and language change over time [17].

This type of work emphasizes the role of (i) the function of language for communication between individuals (*cooperativity*), and (ii) the biophysical constraints of the human body and its environment (*embodiment*) in the explanation for the origin and development of linguistic structure. We are sympathetic to these arguments and share the criticism of a tradition that in some sense equates the *formalisms* of the researcher with the *mechanisms* of the real brain. However, we regret that this general criticism goes hand in hand with a reluctance to use formal models at all. Many researchers have focused instead uniquely on empirical or philosophical approaches (e.g., [17]), or on building "embodied" robots (e.g., [32]).

The goal of this article is to argue that formal models can deal in a meaningful way with embodiment, situatedness, and self-organization. They can help to define these concepts and elucidate the role they play in the development of complex language. *Language games*, such as those studied in recent years in the field of artificial life (see, e.g., [29, 15] for reviews), are a prime candidate for this purpose. Language games are models of language change and language evolution in populations of communicating individuals. Although in most of these models cooperativity and embodiment have not played much of a role, we believe they can be successfully extended to incorporate these important aspects.

The notion of embodiment comes in different flavors. On the one hand, a learning system can be incorporated into an actual robotic body, highlighting the need of the system to cope with sensory limitations [32] and allowing it to manipulate its environment and to develop representations based on sensorimotor interactions with this environment [22]. On the other hand, and more in line with the notion adopted here, embodiment can mean incorporating constraints from sensory, brain, and psychological processing into models without explicitly constructing an artificial body. These approaches are complementary, and neither presents a fully embodied system. In this article we argue that the latter notion of embodiment can be studied with formal models, by incorporating sensory constraints (in the form of noise on the signals) and brain and cognitive processing constraints (by assuming limited processing resources and topological relations between meanings and between signals) into such models.

The models of language evolution that we will consider are *multi-agent models*. They define a population of individuals that talk to each other and learn from each other, using a language that as a result changes over time. Individuals in the models have limited production, memory, and perception abilities, and they have limited access to the knowledge of other individuals. The models evaluate the complex relationship between (i) acoustic, cognitive, and articulatory constraints, (ii) learning and development, (iii) cultural transmission and interaction, (iv) biological evolution, and (v) the complex patterns that are to be explained: the phonology, morphology, syntax, and semantics that are observed in human languages.

The type of language game we examine here is concerned with how a common lexicon can develop in a population of individuals (often called *agents* in this context). In these games, an agent can act either as a speaker or as a hearer. The purpose of a communicative act is the transmission of a meaning from the speaker to the hearer. Meanings cannot be transmitted directly but are encoded by linguistic forms. We can investigate how, based on a great number of such linguistic exchanges under different constraints, a shared lexicon develops so that different speakers use the same word for the same meaning and hearers interpret words with intended meanings. In our models we restrict ourselves to the development of a common lexicon, thus skipping the much more complex and controversial issues in syntax. Nevertheless, we hope

to make the point that language games offer an appealing framework to study other aspects of language as well. For language games that do incorporate grammar, we refer to the extensive review by Kirby [15].

From the perspective of language games, the development of a shared lexicon simply cannot be studied in isolation within one individual, because it depends on the interactions between individuals. In that respect it is a prime example of an aspect of language that escapes study in traditional approaches.

In the rest of this article we will discuss the general framework of these models and present a measure for the quality of a lexicon. We will then study a model that is simple, but is nevertheless novel and serves well to illustrate our approach. Finally, we will discuss how simple language games can be extended to incorporate realistic aspects of cognition, embodiment, and communication.

2 The Optimal Lexicon

The communicative success of a population depends on the organization of the linguistic forms in that population's language, and on how these forms relate to different meanings: how uniquely does one form refer to one meaning? How likely is a speaker to choose a specific linguistic form for a meaning, and how likely is a listener to attribute a certain meaning to a received form? To what extent do individuals agree on the meaning-form mappings? How easily can different forms be confused when communication is noisy?

In this section we will first derive a formal description of what would be the *optimal lexicon*, that is, the lexicon that leads to the highest communicative success in the population. To do so, we need a measure for communicative success. Such a measure is presented next. Similar formalisms were used in [11, 21] and other papers, but our measure is chosen so that we can incorporate some real-world constraints on noise in signaling (like [18]) and different values for different meanings ([14, 19] incorporate in their models the related idea of different frequencies for different meanings).

Speakers can express what they want to say in different ways. Likewise, hearers can interpret spoken forms in different ways. Communicative success is high when the hearer's interpretation of a received form matches with the intention of the speaker. We assume a set of N agents that communicate by forms F to convey meanings M. In a given interaction, a speaker chooses a form f for a meaning m, and the hearer interprets the heard form f^* (which may differ from f if transmission is noisy) and assigns it the meaning m^*. Communication is optimal if speakers and hearers always agree on the meaning for an exchanged form, that is, if $m = m^*$ for any choice of m.

We denote by $S^i(f \mid m)$ the probability that an agent i uses the form f to express the meaning m. Similarly, $R^i(m \mid f)$ is the probability that agent i as a hearer interprets the form f as the meaning m. We assume that there are a finite number $|M|$ of relevant meanings and a finite number $|F|$ of forms used. Further, we assume that similarity between different forms and between different meanings can be measured (e.g., [16]).

We also assume that communication is noisy, that is, the hearer can misperceive a certain form, and more similar forms are more easily confused. We denote by $U(f^* \mid f)$ the probability that an agent perceives the form f as the form f^* (f can be equal to f^*, indicating that the hearer has perceived the form correctly).

Finally, we assume that the communication is successful if the hearer's interpretation is close to the sender's intention. The probability of successfully conveying a certain meaning thus depends on the probabilities of the sender using certain forms and the probabilities of the hearer perceiving and interpreting these forms correctly. We denote by $V(m^*, m)$ the value (or reward) for the hearer understanding m^* when the speaker intended m. Thus V is a measure of communication quality. It should express both

the relative importance of a certain meaning, and the relations between alternative meanings. For example, we could assume that interpreting a signal with a meaning that is wrong but similar is better than interpreting it with just a random meaning, or that being able to express frequent meanings is more important than being able to express infrequent ones.

From these observations, we derive a simple equation that describes the probability $P(m^* \mid m)$ of any hearer j having an interpretation m^* when the speaker i intended m:

$$P(m^* \mid m) = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \sum_{f \in F} \sum_{f^* \in F} \left(S^i(f \mid m) \cdot U(f^* \mid f) \cdot R^j(m^* \mid f^*) \right) \quad (1)$$

This equation says that the probability of the meaning m being perceived as m^* ("understanding m as m^*") is the probability of agent i using the form f to encode meaning m, the hearer perceiving form f^* and then interpreting it as m^*. Because we sum over all N agents as speakers and all but one as hearers ($N-1$; agents do not talk to themselves), we divide the whole expression by $N(N-1)$.

From here it is only a small step to define the communicative success C of the whole population of N agents talking about all $|M|$ meanings:

$$C = \frac{1}{|M|} \sum_{m \in M} \sum_{m^* \in M} \left(P(m^* \mid m) \cdot V(m^*, m) \right) \quad (2)$$

That is, overall communicative success is the sum of the probabilities for all meaning transmissions weighted by their values (assuming that all meanings are equally frequent). This measure is normalized with the number of meanings.

Because S, R, U, and V can all be described as matrices, we can in fact summarize Equations 1 and 2 as follows:

$$C = \frac{1}{|M|N(N-1)} \sum_i \sum_{j \neq i} \left(S^i \times \left(U \times R^j \right) \right) \cdot V \quad (3)$$

where the '$\times$' indicates usual matrix multiplication, and the '$\cdot$' indicates the summation of the product every element in one matrix with its corresponding element in the other matrix (dot multiplication).

Equation 3 constitutes a very general quality measure for a communication system between individuals (described by the matrices S and R), under some *embodied* constraints of articulation and perception (described by U) and semantic/pragmatic constraints on how useful an interpretation is given a certain intention (described by V). By choosing the proper U and V, a wide range of different noise and reward functions can be modeled. However, these matrices can of course not capture all aspects of the embodiment and environment. For instance, the development of conceptual and articulatory abilities and the dependence of rewards and confusion probabilities on specific contexts cannot be modeled directly with our four matrices. However, the formalism is easily extendable to incorporate such aspects. Moreover, even if not all aspects of animal (e.g., [24]), human, or robot communication (e.g., [32]) are modeled, the formalism gives a principled way to abstract out those aspects of embodiment that are nonessential for the emerging language.

With equation 3 in hand, we can now investigate under which conditions communicative success is maximized. We will not provide analytical results for any specific

choice of U and V. Instead, we will present numerical results for a variety of choices of U and V with a simple hill-climbing algorithm. The algorithm used throughout this section is the following:

1. Initialize a population of P individuals, each with an $|M| \times |F|$ matrix S (a production lexicon) and an $|F| \times |M|$ matrix R (a reception lexicon) set with random values and columns normalized.
2. Measure C according to Equation 3.
3. Apply a random change (from a Gaussian distribution with mean 0 and standard deviation $n = 0.1$) to a random entry in a random matrix of a random individual, and normalize the column.
4. Measure C' according to Equation 3.
5. If $C > C'$, revert the change; otherwise $C := C'$.
6. If maximum steps are reached, stop; otherwise go to 3.

Note that in the simulations that use this algorithm an individual's lexicon is not changed as a direct consequence of communication, but is changed randomly. However, this random change may lead to higher communicative success, in which case the change is retained. We use this simple *global optimization procedure* to analyze what the optimal lexicon will look like for different choices of U and V. In Section 3 we will look at the more realistic situation where agents optimize their individual communicative success, that is, where optimization is *local* and *distributed*.

2.1 Categorical Meanings; Noise-Free Signaling

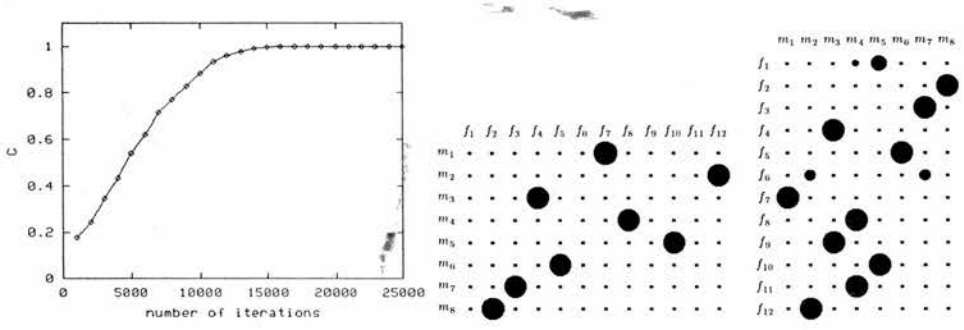
Let us first consider the simplest case of categorical, noise-free communication. That is, we assume that every meaning is unique and has no relation with other meanings. Further we assume that forms are perceived as they are uttered. In short, both U and V are unit matrices (matrices with 1's on the diagonal, and 0's everywhere else).

If we optimize a population's lexicon under these conditions using the hill-climbing algorithm described above, we obtain results as in Figure 1. Here C increases steadily and reaches the optimal value (1.0). The S matrices in the population have maximal probability (= 1.0) for a specific form (horizontal) for each of the meanings (vertical), and probability 0 for all other forms. In the matrix R these forms (vertical) are interpreted as the "correct" meanings. Because there are more possible forms than meanings, some forms are never used and have arbitrary interpretations.

From this simple simulation we can derive two properties of the optimal lexicon: *specificity*, one unique form for every intention, and one unique interpretation for every used form, if $|M| \leq |F|$; and *coherence*, that is, everyone in a population uses the same form for the same meaning.

2.2 Categorical Meanings, Noisy Signaling

If there is noise on the signal (due to a noisy environment and sensory limitations of the hearer), we can expect the hearer to sometimes hear a different form than the speaker uttered. We can model this by introducing nonzero off-diagonal entries in the matrix U. Here, we consider only the simplest case, where forms vary on one axis, determined by their index, and we set the values of U depending on the distance from



(a) Development of communicative success over 25000 iterations

(b) S matrix of a random individual, showing for each meaning (vertical) the probability that she will use any of the forms (horizontal) to

(c) R matrix of the same individual, showing for each form (vertical) the probability that she will choose any of the meanings (horizontal) as its

Figure 1. The optimal lexicon in a population under categorical, noise-free conditions. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are plotted as a small dot. (V and U are unit matrices, $|M| = 8$, $|F| = 12$, $N = 3$, $n = 0.1$).

the "correct" form (and subsequently normalize every row of U):

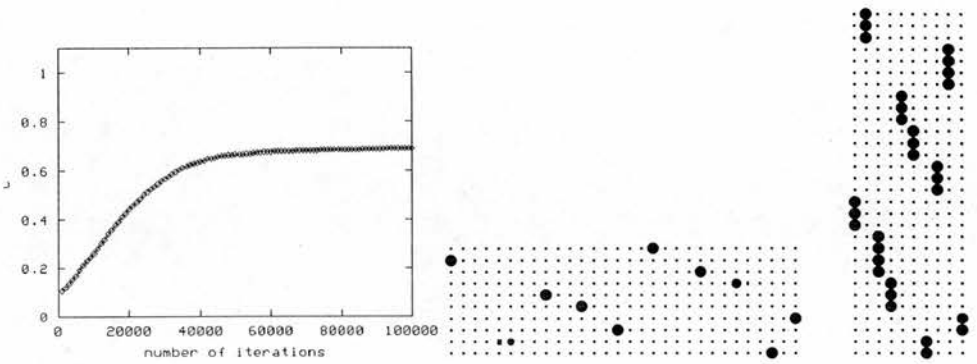
$$U(f^* \mid f) = \frac{1}{1 + (f - f^*)^2} \tag{4}$$

We expect a lower optimal value of C. Moreover, for optimized C, we also expect to find matrices that somehow minimize the chance of misinterpretation. Figure 2 shows that this is indeed what happens. The S matrix shows that for every meaning, there is a prototype form that individuals use. For these prototype forms and their direct neighbors, the interpretation is the "correct" meaning. Thus, little clusters of neighboring forms are all interpreted in the same way, such that prototype forms are maximally distinct from each other. Thus, in addition to specificity and coherence, *distinctiveness* is a property of the optimal lexicon when the signaling is noisy. Note that, even though there are many more forms than meanings, all forms have a specific "best" interpretation. We can obtain similar results with form spaces that have more dimensions [36] or continuous values [37].

2.3 Semantic Similarities and Noisy Signaling

If we include in the model the assumption that not only forms have similarity relations, but also meanings relate to each other, we can identify a fourth criterion of the optimal lexicon: *regularity*. Figure 3 shows results that are obtained by running the hill-climbing algorithm of this section, with U as in Equation 4 and, similarly, V as follows (and rows subsequently normalized):

$$V(m^*, m) = \frac{1}{1 + (m - m^*)^2} \tag{5}$$



(a) Development of communicative success over 100000 iterations

(b) S matrix of a random individual.

(c) R matrix of the same individual.

Figure 2. A local optimum of the lexicon in a population under categorical, noisy conditions (V-unit matrix, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 3$, $n = 0.1$).

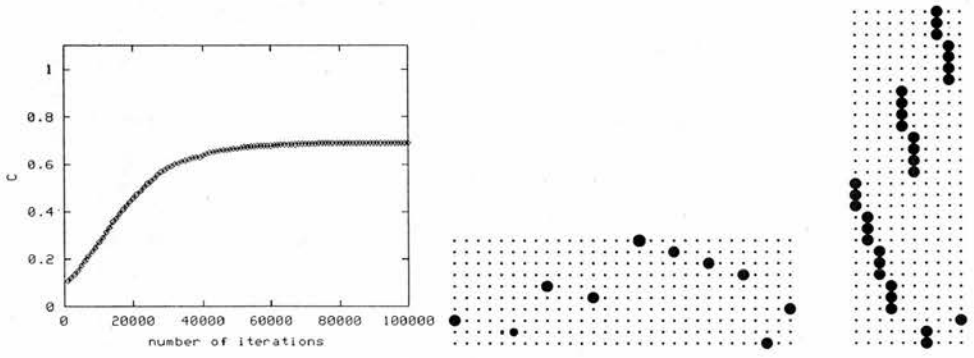
Here, V is maximal when the intended meaning m and understood meaning m^* are the same and decreases with increasing distance between m and m^*.

The local optima found by the hill-climbing algorithm show not only specificity, coherence, and distinctiveness, but also *partial regularity*: similar forms tend to have similar meanings, such that misinterpretations are still better than a random interpretation. The solution found is a local optimum; the globally optimal lexicon is maximally regular: with the parameters of the simulations in Figure 3, meaning m_1 is expressed with form f_1, and forms f_2 to f_3 are interpreted as m_1; meaning m_2 is expressed with f_5, and f_4 to f_6 are interpreted as m_2; and so on. This optimum is not found in this simulation; however, in the local optimum of Figure 3 neighboring clusters of forms are, with only a few exceptions, associated with neighboring meanings. In related work [36] we found that with a slightly different representation the optimum can easily be found as well. Measuring the degree of regularity (as the correlation between the distances between each pair of meanings and the distances between their associated forms) shows that it is consistently higher under conditions with semantic similarities than without.

2.4 Properties of the Optimal Lexicon

From these experiments we can conclude that the optimal lexicon must have the following properties (provided that $|M| \leq |F|$, and that the off-diagonal U and V values are sufficiently low):

- *Specificity*: Every meaning has exactly one form to express it, and vice versa (i.e., there are no homonyms, and no real synonyms: if different forms have the same meaning, they are very similar to each other).
- *Coherence*: All agents agree on which forms to use for which meanings, and vice versa.
- *Distinctiveness*: The forms used are maximally dissimilar to each other, so that they can be easily distinguished.



(a) Development of communicative success over 100000 iterations (b) S matrix of a random individual (c) R matrix of the same individual

Figure 3. Local optima for S and R under semantic similarities, noisy signaling conditions (V as in equation 5, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 3$, $n = 0.1$).

- *Regularity*: In the mapping between meanings and forms there is a *preservation of topology*, that is, similar forms tend to have similar meanings.

3 Language Games

After establishing the properties of an optimal lexicon, we can now turn to *language games*, where there is no global optimization, but rather, every individual tries to optimize its own communicative success. Language game models can be viewed as an extension of the basic communication model that consists of a sender, a message, and a receiver. Language games consider a *population* of individuals (*agents*) that can both send and receive. A language game then is a linguistic interaction between two or more agents that follows a specific protocol and has varying degrees of success. The types of models that we will consider have the following components: (i) a linguistic representation, (ii) an interaction protocol, and (iii) a learning algorithm. In this section we will discuss the choices we have made for each of these components, based on a review of existing models.

3.1 Linguistic Representation

By a *representation* we mean here a formalism to represent the linguistic abilities of agents, ranging from recurrent neural networks [1] or rewriting grammars [13, 35] to a simple associative memory [11, 21, 28, 20, 6, 12, 26], representing the strength of associations between meanings and forms.

In the model of this section, we use the same S and R matrices as in Section 2. Forms and meanings thus remain abstract. Other researchers (e.g., [30, 3]) have chosen more concrete representations, such as random concatenations of consonants and vowels for the forms, or positions in a psychophysically motivated color space for meanings. However, these models do not have similarity relations between forms or between meanings. Instead, forms and/or meanings are categorical, and as a result the form-meaning associations in the emerging languages are completely arbitrary (as in our first model, Sections 2.1 and 2.2). A possible exception is the model in [31]; however, in that article it is not clear whether the stochasticity in the meaning space is dependent on the assumed topology (i.e., whether a wrong but close interpretation is more valuable

than a far-off interpretation), and regularity and distinctiveness are not measured or analyzed.

In contrast to these models, we assume here that there are varying degrees of similarity between forms and between meanings, i.e., there is a topological space of meanings, and a topological space of forms. In that respect, our model is more similar to models of the evolution of grammatical language, where associations between structured meanings and structured forms are not arbitrary (e.g., [14, 2]). For the sake of simplicity, we report here results from simulations where forms and meanings each vary on a one-dimensional axis. As in Section 2, we interpret the index of meanings and forms in the S and R matrices as their positions on these axes. Even such a similarity metric, which is only a first step toward more cognitive plausibility, brings fundamentally new behaviors.

3.2 Interaction Protocol

The agents in language game models interact following simple protocols. In most models two agents—a speaker (initiator) and a hearer (imitator)—are chosen at random. Three types of games can be distinguished. In the *imitation game* [5], in contrast to the present models, meanings play no role. However, as in our model and in contrast to most other language game models, the imitation game assumes noise and similarities in the form space and studies the emergent maximization of the distance between them.

In the imitation game, the initiator chooses a random form from its repertoire and utters it. The imitator then chooses the form from its own repertoire that is closest to the received form and utters it. If the initiator finds that the closest match to this (heard) form is the form that it originally used, the game is successful. Otherwise the game is a failure.

In the *naming game* [28], meanings do play a role. The speaker chooses a meaning and a form to express that meaning, and the hearer makes, based on the perceived form, a guess of what is meant. The hearer then receives feedback from the speaker on the intended meaning, that is, whether its guess was correct. The game is a success if the speaker's intention and the hearer's interpretation are the same, and a failure otherwise. The naming game serves as a model system for studying the emergence of conventional form-meaning associations.

In the *observational game*, the meaning of the expressed form is immediately available to the hearer (as in situations where the speaker points at the object that is the topic of a conversation). This simplification has been used in most language game models studied so far (e.g., [11, 28, 21, 1, 13, 12]).

In the model described here, we make another simplifying assumption. We pick two random agents from the population. The first agent learns from the other, and is randomly assigned the role of either speaker or hearer. We then assume that the first agent is able to assess the *overall* communicative success in communicating with the other agent, and learns through a form of hill climbing as described below. The effect of one interaction in our model can thus be seen as the average effect of many interactions in the naming game. In Section 4 we will discuss the consequences of relaxing this assumption.

3.3 Learning Algorithm

In most models, the learning algorithm that agents use to improve their linguistic abilities is very simple (see [27] for a discussion of the required biases of these learning algorithms and how these biases can evolve). In all of the language game models mentioned above, a mechanism is implemented to keep track of the success of each form or form-meaning association. Whether or not a specific association is used depends on this score. Such algorithms can be considered variants of a hill-climbing process:

given a present state of the system, a random variation is tried out. If the performance is better than before, this variation is kept, and otherwise it is discarded.

The difference from a standard hill-climbing algorithm (such as in Section 2) is that optimization is *local* (every agent optimizes its individual success) and many variants are tried out at the same time. That is, at any one time we can view associations with a high score as constituting the present state of the system. For the other associations, the (low) scores are estimates of how much communication would improve by adopting it. If adopting it would improve communication at this point, the scores will go up and the association will eventually become part of the system.

In the language game model of this section, we will simply use a *local* hill-climbing variant. After picking two random agents, the learning agent makes a random change in its S matrix (if it is assigned the role of speaker) or R matrix (if it is the hearer). The learner checks if that change improves the communicative success in communicating with the other agent according to the following equation (which is almost identical to Equation 3, but now for one specific speaker and hearer):

$$C^{ij} = \frac{1}{|M|}\left(S^i \times \left(U \times R^j\right)\right) \cdot V \tag{6}$$

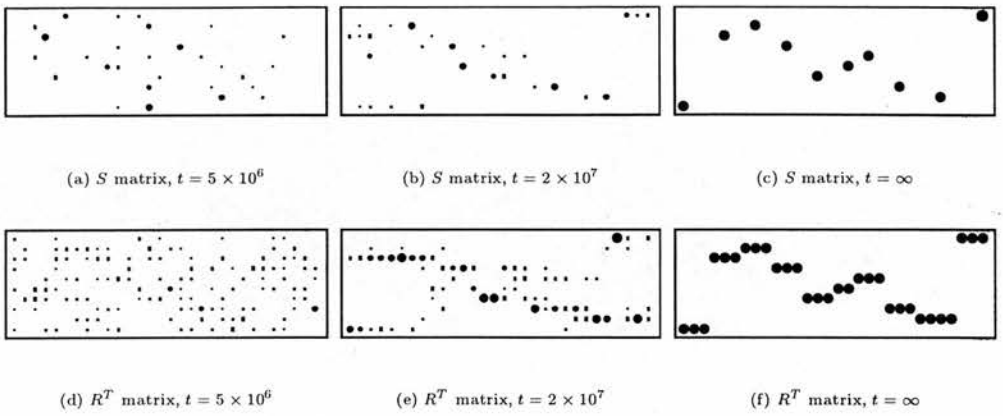
If $C^{ij}_{\text{before}} > C^{ij}_{\text{after}}$, the change is kept; if not, the change is reversed. Note that in this *distributed hill climbing*, at every interaction the target of the hill-climbing process can be different, because each interaction is with a random other agent in the population and because other agents are learning at the same time.

3.4 Self-Organization of the Optimal Lexicon

The main result that we present here is that close approximations of each of the properties of the optimal lexicon emerge from the local interactions that we have defined above. Figure 4 shows results from a simulation with the same parameters as in Figure 3, just with a larger population ($N = 40$) and a higher noise level (the random change in the hill-climbing algorithm is from a Gaussian distribution with mean 0 and standard deviation $n = 1.0$). The figure shows S and R matrices from one random individual at three points in the simulation: after 5×10^6 and 2×10^7 iterations, and in the stable equilibrium configuration (after almost 1×10^8 iterations).

The lexicon that develops shows all four characteristics. In the S matrix at equilibrium (labeled $t = \infty$), every meaning is always expressed by one unique form; in the R matrix, that form is always interpreted with the correct meaning (specificity). At equilibrium, all agents have the same S and R matrices (coherence). In the S matrix, the total distance between all preferred forms is (almost) maximal; in the R matrix, each of these preferred forms (except at the edges) is the center of a little cluster of forms that are all interpreted with the same meaning (distinctiveness). Finally, with three exceptions, all form clusters have neighboring form clusters that express a neighboring meaning (regularity).

The degree of regularity in this simulation is small (the correlation between the distance between each pair of meanings and the distance between their corresponding forms is around 0.2). In general, regularity can be difficult to obtain because to go from an irregular to a regular lexicon many changes to the lexicon are required. Moreover, its contribution to the communicative success is small in comparison with the other three properties. In [36] we show results with a different representation, where the entries in the S and R matrices are always 1 or 0, and random changes move a 1 to a different position in the matrix. In this setup regularity can much more easily emerge, both in the global and in the distributed hill-climbing condition.



(a) S matrix, $t = 5 \times 10^6$ (b) S matrix, $t = 2 \times 10^7$ (c) S matrix, $t = \infty$

(d) R^T matrix, $t = 5 \times 10^6$ (e) R^T matrix, $t = 2 \times 10^7$ (f) R^T matrix, $t = \infty$

Figure 4. Development of specificity, coherence, distinctiveness, and regularity in the lexicon of a population under semantic similarities, noisy signaling conditions. At each time step a random speaker interacts with a random hearer and one of them performs a single hill-climbing step to improve the communication. In this graph, the R matrices are transposed, so that in both S and R^T meanings are on the vertical axis and forms on the horizontal axis. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are not plotted. $t = \infty$ indicates any time after the simulation has converged (from around $t = 10^8$) to the stable equilibrium. (V as in Equation 5, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 40$, $n = 1.0$.)

4 Toward More Cognitive Plausibility

Our results show that there is no necessity for explicit and innately specified "principles" that guarantee specificity, distinctiveness, coherence, and regularity. It is possible in principle that these basic characteristics emerge from simple interactions between agents, a generic learning algorithm, and topological meaning and form spaces. That is, they emerge from the embodiment (i.e., general perceptual and processing constraints) and situatedness (i.e., interactions between individuals) of the simulated agents.

Of course, the biophysical constraints of real humans are different from the ones implemented in this model. The next step in our research is therefore to evaluate whether more *realistic* constraints lead—through similar dynamics—to an emergent language with more *realistic* characteristics. Here we consider three possible extensions of the model.

4.1 Limited Feedback

In the distributed hill-climbing simulations we assumed that an agent makes a random change in one of its matrices, and then evaluates if that change increases the success in communicating with one other individual. In reality, that information might not be available. It is therefore worth examining if the same results can be obtained with the minimal assumptions of feedback on whether or not a communication about a single meaning has been successful (as in the naming game [28]), or on shared contexts between speaker and hearer (as in the observational game [25]).

We have done some experiments that show that at least specificity, coherence, and distinctiveness can easily emerge in a naming game setup [37]. Figure 5 shows one of the emerging languages from these experiments. It shows a pattern formed through local interactions between two communicating agents, expressing nine different meanings with forms from a two-dimensional form space. Each of the nine clusters in this figure shows strong associations from two agents for one particular meaning.

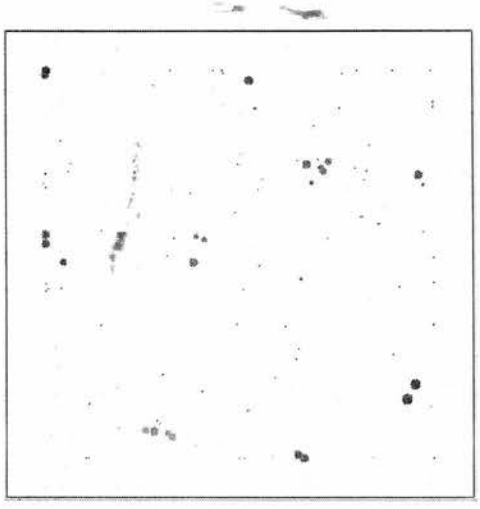


Figure 5. Local interactions: emergence of distinctiveness, coherence, and specificity. Dispersed forms in form space, obtained through local interactions between communicating agents. Each of the nine clusters in this figure shows associations from both agents for one particular meaning. Large dots are strong association. (Parameters: $N = 2$, $|M| = 9$; form space continuous—i.e., $|F| = \infty$; perceptual noise 10%.)

4.2 Cooperativity

An important principle in line with the joint-action view of human communication has been formulated by Grice [9] as the *principle of cooperation*: In a conversation, the speaker makes certain assumptions about the expectations of the hearer, and she uses these assumptions to communicate her intended message effectively. This principle involves the provision of enough, but not too much, information in a message, the relevance of the message to the current conversation topic, and the truthfulness of the information provided. In interpreting the message, the hearer relies on the speaker to have obeyed these principles.

In the context of language game models, we can extend this principle to the cooperative creation of new words: a speaker that is interested in communicative success should only generate a new form if no form for the intended meaning already exists in the language. For example, a speaker who wants to talk about a duck-billed platypus but has forgotten the name for it (or never knew it) would not make up a random word and thus confuse the hearer. Instead, she would either circumvent the term or describe the animal, and somehow prompt the hearer to give the name. By querying the hearer for a possible form, the speaker allows herself to make assumptions about the beliefs of the hearer and therefore to engage in a *cooperative* language game (as opposed to the merely *interactive* language games that are traditionally studied). Such an extension of the language game framework is plausible in that it views language as a cooperative phenomenon and as a means to maximize the efficiency of communicating intended meanings. It will prevent the creation of an excess of new forms, thereby reducing the number of synonyms and the cognitive load.

4.3 Analogy

When an agent creates a new form in a language game, it usually randomly assembles phonemes (e.g., [28]). This mechanism is in line with the claim of the "arbitrariness of the sign" [7]: the structure of the form has no relationship to the meaning conveyed by it. While this is true for many forms in today's existing languages, there is evidence suggesting that, in the creation of new forms, the intended meaning should be taken

into account. First of all, when new words are created in, for example, English, they are often compounded and derived from existing words to ease their understanding. Thus, someone who eats bananas will be called a "banana-eater" rather than a "manslo," to indicate the semantic relationship with bananas and eaters. While such a process cannot be applied to simple language games directly, it does show a structural relationship between words that reflects a semantic relationship between their meanings.

Second, there is growing evidence for the hypothesis that the sound of a word can suggest its meaning ("sound symbolism"). This idea was first mentioned by Plato and has been pursued since then, for example, by von Humboldt [34]. Subsequent psycholinguistic research has shown that in the formation of words, certain sounds can represent certain meanings. For example, in assigning the two words *Mil* and *Mal* to images of big and small tables, 80% of subjects chose *Mal* to stand for the larger table and *Mil* for the smaller table, indicating that /a/ suggests large size and /i/ small size [23]. These results have been reproduced and extended by numerous researchers (see e.g. [10]).

A less controversial version than such *absolute* sound symbolism (where sounds carry meaning) is a *relative* sound symbolism that can be directly applied to the creation of new forms in naming games. It is described by von Humboldt [34, p. 74] as "Words whose meanings lie close to one another, are likewise accorded similar sounds," while the sounds themselves bear no direct semantic content. In Sections 2 and 3 we presented results where such relative sound symbolism (regularity) emerges as an optimal solution in noisy conditions. However, we can also imagine that agents actively exploit a form of topology preservation when creating new forms. In a language game the decoding of the form by the hearer could then work as follows:

```
Find a meaning for the form f:
for the nearest neighbor f' of f according to the similarity
    metric, find the best meaning m'
associate f with a meaning which is closest to m'
```

This approach can help to reduce ambiguity in the hearer's lexicon. Preliminary results suggest faster convergence of the language than in the original model, due to the emergence of regularities in the form-meaning mapping. Further, we found several examples of parameter settings that would not lead to convergence under the classical settings, but did converge under these topological settings. Finally, we find an unexpected delay in the convergence in the final stage, due to conflicts between competing partial regularities. This delay indicates that lexicon creation is subject to the opposing pressures of *topological preservation* and *distinctiveness maximization*. We could assume that in the evolution of vocabularies of human languages, words with similar meanings might have developed to be as similar as possible (and thus predictive of their meaning) while at the same time being as distinctive as possible (to facilitate communication with already known words). A new form that is created to be similar to another in order to facilitate understanding of its meaning would then undergo variation (historical change) to become more arbitrary as it became more established and a prediction of its meaning became less important than its distinctiveness from other forms. While we have not incorporated these constraints in our current simulations, we believe that they present a promising direction in the endeavor to integrate language game formalisms with cognitive approaches to language.

5 Conclusions

We have discussed the relevance of language evolution models to the study of embodiment and self-organization of language, and presented a formalism for describing

language games. Language game models are complementary to work that studies language processing and language acquisition. The models we discussed are simple; their value is that they make the roles of diachrony, embodiment, and self-organization in emerging linguistic structure explicit and testable.

We have argued that the environment and embodiment of communicating agents in the real world impose a topology on both the meaning and the form space of their communication system. We have shown that with these topologies the optimal lexicon has four characteristics: specificity, coherence, distinctiveness, and regularity. We have further shown that in a distributed population of agents that each have generic learning capabilities, a lexicon can be established that shows each of these four characteristics.

Our results on distinctiveness and regularity follow naturally from the framework that we have described in this article. Nevertheless, they have not been reported in the extensive literature on the modeling of language evolution. We believe that this fact in itself is support for our approach to embodiment, where we try to incorporate constraints from sensory, brain, and psychological processing into formal models without explicitly constructing an artificial body. However, much work remains to be done on explaining the role of these constraints in the evolution of language. In the final part of the article, we have therefore raised issues where cognitive science can inform language game modeling, and eventually lead to a detailed understanding of how complex language has emerged from many simple interactions.

Acknowledgments

The writing of this article was supported by European Commission RTN grant HPRN-CT-2000-00065 to Gert Westermann, and a Prins Bernhard Cultuurfondsbeurs and a Marie Curie fellowship of the European Commission to Willem Zuidema. Part of the research described in this article was performed while WZ was at the A.I. Lab of the Vrije Universiteit Brussel and funded through a Concerted Research Action fund (G.O.A.) of the Flemish Government and the VUB.

We thank Kenny Smith, Charlotte Hemelrijk, Hanspeter Kunz, and two anonymous reviewers for helpful comments.

References

1. Batali, J. (1998). Computational simulations of the emergence of grammar. In J. Hurford & M. Studdert-Kennedy (Eds.), *Approaches to the evolution of language: Social and cognitive bases.* Cambridge, UK: Cambridge University Press.
2. Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge, UK: Cambridge University Press.
3. Belpaeme, T. (2001). Simulating the formation of color categories. In B. Nebel (Ed.), *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01)* (pp. 393–398). San Francisco: Morgan Kaufmann.
4. Clark, H. H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.
5. De Boer, B. (1999). *Self-organisation in vowel systems.* Ph.D. thesis, Vrije Universiteit Brussel AI lab.
6. de Boer, B., & Vogt, P. (1999). Emergence of speech sounds in changing populations. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), *Advances in Artifical Life* (pp. 664–673). Berlin: Springer-Verlag.
7. de Saussure, F. (1916). *Course in general linguistics.* La Salle, IL: Open Court. Translated by Roy Harris. Edition published in 1986.

8. Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisin, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

9. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (pp. 41–58). New York: Academic Press.

10. Hinton, L., Nichols, J., & Ohala, J. J. (Eds.) (1995). *Sound symbolism.* Cambridge, UK: Cambridge University Press.

11. Hurford, J. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua, 77,* 187–222.

12. Kaplan, F. (2000). *L'émergence d'un lexique dans une population d'agents autonome.* Ph.D. thesis, Université Paris 6, Sony CSL-Paris.

13. Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, J. Hurford, & M. Studdert-Kennedy (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form.* Cambridge, UK: Cambridge University Press.

14. Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5,* 102–110.

15. Kirby, S. (2002). Natural language from Artificial Life. *Artificial Life, 8,* 185–215.

16. Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

17. MacWhinney, B. (Ed.) (1999). *The emergence of language.* Mahwah, NJ: Lawrence Erlbaum Associates.

18. Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences of the U.S.A., 96,* 8028–8033.

19. Nowak, M. A., Plotkin, J. B., & Jansen, V. A. (2000). The evolution of syntactic communication. *Nature, 404,* 495–498.

20. Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior, 7.*

21. Oliphant, M. & Batali, J. (1996). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter, 11.*

22. Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

23. Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12,* 225–239.

24. Seyfarth, R. M., & Cheney, D. L. (1997). Some general features of vocal development in nonhuman primates. In C. T. Snowdon & M. Hausberger (Eds.), *Social influences on vocal development* (pp. 249–273). Cambridge, UK: Cambridge University Press.

25. Smith, A. D. (2001). Establishing communication systems without explicit meaning transmission. In J. Kelemen & P. Sosík (Eds.), *Advances in Artificial Life (Proceedings 6th European Conference on Artificial Life, Prague).* Berlin: Springer.

26. Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science, 14,* 65–84.

27. Smith, K. (2003). Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior, 10*(1), 25–44.

28. Steels, L. (1997). Self-organising vocabularies. In C. Langton & K. Shimohara (Eds.), *Proceedings of the 5th International Workshop on Artificial Life: Synthesis and simulation of living system* (pp. 179–184). Cambridge, MA: MIT Press.

29. Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication, 1,* 1–35.

30. Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence, 103*, 133–156.

31. Steels, L., & Kaplan, F. (1998). Stochasticity as a source of innovation in language games. In C. Adami, R. Belew, H. Kitano, & C. Taylor (Eds.), *Proceedings of Artificial Life VI* (pp. 368–376). Cambridge, MA: MIT Press.

32. Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The Transition to Language*. Oxford, UK: Oxford University Press.

33. Tomasello, M. (Ed.) (1998). *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Lawrence Erlbaum Associates.

34. von Humboldt, W. (1836). *On Language*. Cambridge, UK: Cambridge University Press. Translated from the German by Peter Heath. Edition published in 1988.

35. Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15 (proceedings of NIPS'02)* (pp. 51–58). Cambridge, MA: MIT Press.

36. Zuidema, W. (2003). Optimal communication in a noisy and heterogeneous environment. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, & J. Ziegler (Eds.), *Advances in Artificial Life (proceedings of the 7th European Conference on Artificial Life)* (pp. 553–563). Berlin: Springer Verlag.

37. Zuidema, W. & Westermann, G. (2001). Towards formal models of embodiment and self-organization of language. In *Proceedings of the Workshop on Developmental Embodied Cognition*. Edinburgh, UK.

nodes, there are 2^n linear orders. This can quickly lead to large numbers, but these are still smaller than the $m(!)$ possible permutations that would result from m constituents (for $m = 3$ as in the present example, this gives six).

The present proposal is that syntax does indeed provide only for the more modest constraints given by a-temporal syntax. A-temporal syntax is sufficient to specify a crucial ingredient of syntax, called structure-dependence in many of Chomsky's publications. Structure dependence is decidedly not the specification of linear order, but the specification of domination and sisterhood alone.

Order of constituents is only partially determined by structure dependence. The remaining task is that of phonology, semantics, and pragmatics combined. I have nothing to say about the latter two, but will assume that principles of information structure (such as "Agent First" and "Focus Last," *Foundations*, Ch. 8, sect. 8.7) are of primary importance here. Again, avoidance of duplication seems to make a syntactic determination of order superfluous at best in those cases in which other principles are at work already.

4. The role of phonology. As for linear order in phonology, it is indisputable that phonology (in contrast to syntax) needs linear order as a core concept. The string of phonemes /pit/ is in contrast with the string /tip/, while /ipt/ is a possible, but unrealized word in English, and any other permutation of the three phonemes is ill-formed in English. In other words, the elementary notions of contrast, distinctiveness, and well-formedness in phonology include linear order. Structuralist phonology used the term "syntagmatic relation" in this connection; here, "syntagmatic" literally means "in accordance to the time axis." Furthermore, a number of phonological rules are generally cast in terms of linear order. For example, the basic rule of compound stress in English or German says that the *first* of two parts in a compound carries main stress. For stress in phrases, the reverse holds (simplifying considerably): the *second* of two constituents in a phrase receives main stress. In other words, phonology is very much about the temporal line-up of chunks of speech. Given that it is grounded in the phonetics of speech, this does not come as a surprise.

Furthermore, some of the syntactic movement operations assumed in syntactic theory are clearly related, at least functionally, to either information structure (as in "topic first") or to preferred positions for constituents with either strong stress (focus positions) or weak stress (deaccentuation). Given that syntax is not conceived as "knowing" about nonsyntactic principles such as stress, it is almost inevitable to assign the respective movement operations to some other domain.

5. Where does order come from? If the present hypothesis about temporally unordered syntactic constituents should be correct, it would leave us with one crucial question: From what rules or principles does the actual order (encoded in phonological structures) derive? No complete answer can possibly be given here, but parts of the answer have been identified already: Jackendoff points out in several places that there are principles of ordering which are part of semantics, information structure in particular, and of phonology, heaviness constraints and stress preferences in particular.

Lexical information (either on individual items or on more or less extended lexical classes) must be another source of temporal order: Prepositions versus postpositions are an obvious example, prenominal versus postnominal adjectives might provide a further case.

Next, phonology itself provides ordering information, as we can see from principles, such as the one requiring long constituents to follow short ones (Behaghel's law).

Setting aside the cases just enumerated, there are substantial remaining problems. My formal proposal at this point is that the rules providing the interface between syntax and phonology – Jackendoff's "PS-SS interface rules" (Ch. 5, sect. 5.6) – provide the natural locus for stating the constraints on linear order for syntactic and/or semantic constituents. Such rules are, by necessity, sensitive to information stemming from both of the components between which they mediate. Here again, the architecture of grammar proposed by Jackendoff provides a fruitful base for further research.

How did we get from there to here in the evolution of language?

Willem Zuidema[a] and Bart de Boer[b]

[a]*Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, and Institute of Animal, Cell and Population Biology, University of Edinburgh, Edinburgh EH8 9LL, United Kingdom;* [b]*Kunstmatige Intelligentie, Rijksuniversiteit Groningen, 9712 TS Groningen, The Netherlands.* **jelle@ling.ed.ac.uk** **b.de.boer@ai.rug.nl** **http://www.ling.ed.ac.uk/~jelle** **http://www.ai.rug.nl/~bart**

Abstract: Jackendoff's scenario of the evolution of language is a major contribution towards a more rigorous theory of the origins of language, because it is theoretically constrained by a testable theory of modern language. However, the theoretical constraints from evolutionary theory are not really recognized in his work. We hope that Jackendoff's lead will be followed by intensive cooperation between linguistic theorists and evolutionary modellers.

There has been a vigorous debate in the evolution of language literature on whether the human capacity for language evolved gradually or with an abrupt "big bang." One of the arguments in favor of the latter position has been that human language is an all or nothing phenomenon that is of no value when only part of its apparatus is in place. From a developmental perspective this has always been a peculiar argument, seemingly at odds with the gradual development of phonological, syntactic, and semantic skills of infants. In the context of the evolution of language, the argument was eloquently refuted in a seminal paper by Pinker and Bloom (1990). However, Pinker and Bloom did not go much further than stating that a gradual evolution of Universal Grammar was possible. They did not explore the consequences of such a view for linguistic theory, and their approach was criticized by both the orthodox generativists and the latter's long-term opponents.

Jackendoff (2002) has now gone one step further. If linguistic theory is incompatible with gradual evolution and development, perhaps linguistic theory needs to be revised. Jackendoff has written a powerful book around the thesis that the language capacity is a collection of skills ("a toolbox"). Some of these skills are language-specific, some not, and each of them is functional even without all or some of the other skills present. From his decomposition of linguistic skills follow a number of hypotheses on plausible intermediate stages in the evolution of language, that fit in neatly with many other theories, models, and findings in this field.

Jackendoff's book therefore presents a significant departure from the generative, "formalist" tradition, where the evolution of language has received little attention. In this tradition, the structure of human language has often been viewed as accidental rather than as adapted to the functions that language fulfills in life. Chomsky and others have been dismissive about attempts to reconstruct the evolution of language, which they regard as unscientific speculation. Chomsky famously observed that "we know very little about what happens when 10^{10} neurons are crammed into something the size of a basketball" (Chomsky 1975).

In contrast, Jackendoff presents the different tools from the "toolbox" as adaptations for better communication. Moreover, he gives a rather complete scenario of successive, incremental adaptations that is consistent with his view on how modern language works, and how it can be decomposed. Interestingly, he argues that present-day languages show "fossils" of each of the earlier stages: expressions and constructions that do not exploit the full combinatorial apparatus of modern language. Jackendoff's book is therefore a major contribution towards a more rigorous, scientific theory of the evolution of language, in part because it leads to some testable predictions, but more importantly because it is theoretically constrained by a testable theory of modern language.

However, Jackendoff does not really recognize that, in addition, evolutionary theory brings stringent theoretical constraints (Barton & Partridge 2000). Good evolutionary explanations specify the assumptions on genotypic and phenotypic variation and selection pressures, of which the consequences can be worked out in mathematical and computational models. For instance, Nowak et al. (2001) derive a "coherence threshold" for the evolution of language, which poses a strict constraint on the accuracy of both genetic and cultural transmission of language for linguistic coherence in a population to be possible. In this type of work, one often finds that "adaptive explanations" that seem so obvious in a verbal treatment such as Jackendoff's, are in fact insufficient.

Cavalli-Sforza and Feldman (1983) studied a "conformism constraint" that arises from the positive frequency dependency of language evolution: Linguistic innovations are not advantageous in a population where that innovation is very infrequent. Imagine, for instance, a population that is in the second state of Jackendoff's scenario. That is, individuals can use a large vocabulary of learned signals in a non-situation-specific manner, but their language is not compositional: Signals cannot be analyzed as consisting of meaningful parts. Suppose that a child is born with a genetic mutation that makes her more inclined to analyze sentences compositionally. Would this child profit significantly from this mutation, even if the language of the population she is born into is not at all compositional? If not – and it takes some creativity to come up with reasons why she would – evolutionary theory predicts that the new gene will disappear through negative selection or random drift (Fisher 1922).

That is not to say that language did not evolve according to Jackendoff's scenario, but just to emphasize that each of the transitions between the phases he proposes is a challenge in itself. The evolution of language is not, as is sometimes suggested, a domain for just-so stories. Rather, it turns out that it is very difficult to find even a single plausible scenario for the evolutionary path from primate-like communication to the sophisticated toolbox of human language that will survive close scrutiny from mathematical and computational modeling. Recently, this insight has led to a surge in the interest in "explorative," computational models (see Kirby 2002b; Steels 1997; for reviews). They have yielded intriguing ideas on adaptive and nonadaptive explanations for the emergence of shared, symbolic vocabularies (e.g., Oliphant & Batali 1996), combinatorial phonology (e.g., de Boer 2000; Oudeyer 2002), compositionality and recursive phrase-structure (e.g., Batali 2002; Kirby 2002a).

For instance, the suggestion of Kirby (2000) – referred to but not discussed in Jackendoff's book – is that a process of cultural evolution might facilitate the emergence of compositionality. If a language is transmitted culturally from generation to generation, signals might frequently get lost through a bottleneck effect (that arises from the finite number of learning opportunities for the child). Signals that can be inferred from other signals in the language, because they follow some or other systematicity, have an inherent advantage over signals that compete for transmission through the bottleneck. With some sort of generalization mechanism in place (not necessarily adapted for language), one always expects a language to become more compositional (Kirby 2000), and, more generally, better adapted to the idiosyncrasies of the individual learning skills (Zuidema 2003).

Throughout his book, Jackendoff uses metaphors and terminology from computer science. Terms like processing, working memory, and interface make it sometimes appear as if he is describing a computer rather than processes in the human brain. However, nowhere do his descriptions become sufficiently formal and exact to make them really implementable as a computer program. In this light, his criticism of neural network models of language acquisition and his mentioning only in passing of computational models of the evolution of language is unsatisfactory. Jackendoff's challenges for connectionists are interesting and to the point, but it is equally necessary for theories such as Jackendoff's, especially their implications for development and evolution, to be made more precise and to be extended in computational and mathematical models.

In sum, in the effort to find a plausible scenario for the evolution of human language, a book like Jackendoff's *Foundations of Language*, based on a broad and thorough review of linguistic theory and facts, is extremely welcome. But as explorative computational models such as the ones discussed have been very fruitful in showing new opportunities and constraints for evolutionary explanations of human language, we hope that Jackendoff's lead will be followed by intensive cooperation between linguistic theorists and evolutionary modellers.

ACKNOWLEDGMENT

Willem Zuidema is funded by a Marie Curie fellowship of the European Union.

Author's Response

Toward better mutual understanding

Ray Jackendoff

Program in Linguistics, Brandeis University, Waltham, MA 02454.
jackendoff@brandeis.edu

Abstract. The commentaries show the wide variety of incommensurable viewpoints on language that *Foundations of Language* attempts to integrate. In order to achieve a more comprehensive framework that preserves genuine insights coming from all sides, everyone will have to give a little.

R1. Goals

My goal in writing *Foundations of Language* was threefold. First, I wished to develop a framework for studying language – the parallel architecture – which would permit a better integration of all the subfields and theoretical frameworks of linguistics with each other and with the other cognitive neurosciences. Second, I wished to persuade linguists to join more fully in this integrative enterprise. Third, I wished to persuade cognitive neuroscientists outside linguistics that the past forty years have brought genuine insights in linguistic description – albeit somewhat obscured by the technical opacity of linguistic theory – and that the parallel architecture offers better prospects for renewed dialogue. The commentaries suggest that I have succeeded to some extent, but that there still is a long way to go and a lot of preconceptions to overcome (including, no doubt, my own). The difficulties of integration are legion: The study of language, more than any other cognitive capacity, stretches the limits of interdisciplinarity, all the way from neuroscience and genetics to social policy and literary theory, with linguistics, psychology, and anthropology in between.

Many of the commentators focus on issues in *Foundations* that are touched upon only tangentially or not at all in the précis appearing here. In this response I will do my best to make clear what is at stake. My hope, of course, is that readers will thereby be engaged enough to want to tackle the whole book.

Evolution: The erratic path towards complexity

Nick Barton* and Willem Zuidema*†
*Institute of Animal, Cell and Population Biology
†Language Evolution and Computation Research Unit
University of Edinburgh

June 19, 2003

A central goal of evolutionary biology is to explain the origin of complex organs — the ribosomal machinery that translates the genetic code, the immune system that accurately distinguishes self from non-self, eyes that can resolve precise images, and so on. Although we understand in broad outline how such extraordinary systems can evolve by natural selection, we know very little about the actual steps involved, and can hardly begin to answer general questions about the evolution of complexity. For example, how much time is required for some particular structure to evolve? In a recent paper, Lenski et al. [1] give an intriguing example of how "digital organisms" can evolve. Their work suggests many lines of research, which might shed new light on an old problem.

Complex systems — systems whose function requires many interdependent parts — are vanishingly unlikely to arise purely by chance. Darwin's explanation of their origin is that natural selection establishes a series of variants, each of which increases fitness. This is an efficient way of sifting through an enormous number of possibilities, provided there is a sequence of ever-increasing fitness that leads to the desired feature. To use Sewall Wright's metaphor, there must be a path uphill on the "adaptive landscape".

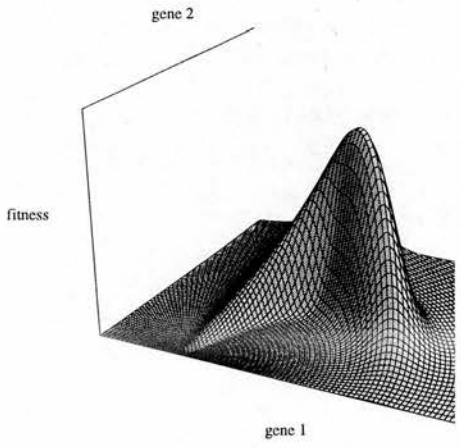


Figure 1: The adaptive landscape (usually) is a graph of fitness against genotype. Sketched is a hypothetical example, in which alleles at two genes have a continuous range of effects. Both real and digital organisms have, in contrast, a discrete set of possible genotypes involving many more than two genes. Thus, mutations can take them in very many directions. This high dimensionality makes it more likely that there is some path uphill to the "adaptive peak" (See [2])

The crucial issue, then, is to know what variants are available (what can be reached from where) and what is the fitness of these variants. Is there a route by which fitness can keep increasing? Population genetics is not much help here. Given the geometry defined by mutation and recombination, and given the fitnesses, we can work out how a population will change, simply by following the proportion of different types through time. But understanding

how complex features evolve requires plausible models for the geometry of the adaptive landscape, which population genetics by itself does not provide.

"Artificial Life" — the study of life as it could be — provides a variety of such models. For instance, Thomas Ray [3] developed a model, called "Tierra", where digital creatures are little computer programs that copy themselves and compete with each other for memory and processing time. Fitness here — just as in the real world — is defined very indirectly by the rate of self-replication of the creatures relative to others. Ray's creatures evolved strategies to hinder competitors and even to parasitize other creatures. Karl Sims [4] created a simulated physical world in which "digital creatures" successfully evolve both their bodies and brains in order to beat other creatures in a variety of tasks such as swimming, walking and jumping. Lipson and Pollack [5], in a recent follow-up study, actually made such walking creatures as little robots and showed that the evolved locomotion strategies work even in the real world. Fitness in these models is defined implicitly by the complex relation between brain and body architecture and the resulting way of moving.

In Lenski et al.'s recent study, the creatures consist of a string of instructions, each instruction being chosen from 26 possibilities. Like Ray's creatures, the instructions must implement self-replication in order for the creature to have offspring. But like Simm's creatures, they are also rewarded for performing a specific task: they can replicate faster by manipulating information from the environment. Each organism receives two random 32 bit strings as inputs, and is rewarded if it produces an output string that matches one of 9 possible logical operations. For example, the logical operation NAND ("not and") returns a 0 in the output string only if the corresponding digits in the input strings are both 1, and a 1 in all other cases. One of the 26 possible instructions in a creature's "genome" is a logic operation (NAND), whilst the others perform various manipulations: copying, input/output, etc. Composite logic operations are valued according to the number of elementary NAND operations needed to perform them. The most valuable is EQU ("equal"), which returns a 1 only if both input bits are the same (this requires 5 NAND operations) as well as other operations with move intermediate results between registers. A hand-written program required 19 operations to achieve EQU; a digital organism needs additional code for replication.

Initially, 3600 identical organisms were set up, each with 15 instructions that allowed replication, plus 35 dummy instructions. In each replication, point mutations occured at a rate 0.0025 per instruction, and single-instruction insertions or deletions occurred at a rate of 0.056 per genome. In one run, EQU evolved after 111 steps. (A "step" is counted whenever offspring differed from parent along the successful lineage; in most cases steps corresponded to single mutations, but 8 steps involved two or three mutations). Over a further 233 steps, the ability to perform additional logic operations evolved, and so fitness increased further. The way in which these organisms evolved was broadly as one would expect. In particular, the evolution of EQU depends on there being fit steps that lead up to it, as allowed by the reward system shown in Table 1.

Function name	Logic operation	Reward
NOT	$\neg A$	2
NAND	$\neg(A \textit{ and } B)$	2
AND	$A \textit{ and } B$	4
OR_N	$(A \textit{ or } \neg B)$	4
OR	$(A \textit{ or } B)$	8
AND_N	$(A \textit{ and } \neg B)$	8
NOR	$\neg A \textit{ and } \neg B$	16
XOR	$(A \textit{ and } \neg B) \textit{ or } (\neg A \textit{ and } B)$	16
EQU	$(A \textit{ and } B) \textit{ or } (\neg A \textit{ and } \neg B)$	32

Table 1: Rewards for performing logical operations. The symbol $\neg$ denotes negation ("not"). Logic operations are performed digit by digit on one or two input strings. Thus, when applied to the input strings "110" and "011", the operation AND would yield "010".

Lenski et al. experimented with the computer model in much the same way that geneticists experiment with model organisms, by changing the fitness regimes and by knocking out instructions on the evolved genomes one at a time to test their effect on fitness. They can also do something that geneticists usually cannot: trace back the evolutionary history of the genome that first produced EQU. From the results from this study, Lenski et al. emphasise one feature in particular: often, deleterious changes are established along the path to evolution of EQU. From a population genetics point of view this result is less surprising than it may seem at first sight. One expects some deleterious mutations to

be picked up by random drift in a population of only 3600 organisms. Moreover, these digital organisms are asexual, so that a deleterious mutation can be established if it occurs together with a favourable mutation (hitch-hiking, [6, 7]), or if a new mutation occurs that produces a fit genotype when *combined* with the initially deleterious mutation. In the example analysed by Lenski et al., most of the deleterious mutations along the lineage leading to EQU only reduced fitness slightly, by less that 3%. However, two reduced fitness by more than 50%, and were only rescued by mutations which occurred immediately afterwards — in one case, by the mutation which first produced EQU. Moreover, that evolution of EQU *required* the previous mutation, which initially greatly reduced fitness. This pattern, of strong epistatic interaction, was seen in the final stages of 3 of the 23 replicates in which EQU evolved.

So, in these simulations adaptation frequently depends on the occurrence of double mutations, either in the same generation, or in close succession. Suppose that a particular deleterious mutation arises at rate μ_1 and reduces fitness by s. It is expected to persist for $\sim 1/s$ generations [8] during which time mutations at another locus occur at rate μ_2. If both occur together, they confer a strong advantage, and are picked up by selection. So, we expect a rate of accumulating these interacting pairs of $\sim \mu_1\mu_2/s$, compared with $\sim \mu_1$ for single favourable mutations. The observation that interacting pairs do get established quite frequently tells us something about the relative abundance of paths involving single mutations versus double mutations: possibly, once all single-mutation steps have been explored, the population must wait until the rarer doublets arise.

In Lenski et al.'s artificial organisms, the mutation rate per site is quite high (0.0025) so that favourable pairs can be picked up by selection at an appreciable rate; this would be unlikely in most real organisms, because there, mutation rates at each locus are low. There are, however, some biological examples in which double mutations contribute to adaptation — the first deleterious, the second favourable in combination. In general terms, Manfred Eigen has argued that evolving populations of RNA molecules form a "quasi-species" [9], with high diversity maintained by predominantly deleterious mutation away from a wild-type sequence that is itself vanishingly rare. This diversity allows the population to explore a larger fraction of sequence space. More specifically, the secondary structure of rRNA molecules can be determined through the pattern of covariation of substitutions: if one base changes, its partner changes soon after in order to maintain base pairing. Here again, the first change occurs by chance, in opposition to selection, and is compensated by the second [10]. Lenski et al. do not explore the applicability of their model to such issues.

Artificial Life models such as Lenski et al.'s are perhaps interesting in themselves, but as biologists we are concerned here with the question what Artificial Life can tell us about real organisms. The difficulty in answering that question is that much work in this field is rather isolated from traditional evolutionary biology. Well-established theories and methods from population genetics and game theory are too often ignored — and Lenski et al., although they explore the evolutionary dynamics in quite detail, are no exception. There are, however, ways in which Artificial Life can benefit from evolutionary theory, and vice versa. Can we understand exactly how complexity evolves in these artificial models? Can we find general rules which describe the process? For example, could we predict how long it is likely to take for a function such as EQU to evolve, given mutation rates and fitnesses? Here, there are population genetics principles which are helpful: the relative rates of single vs. double mutations that we discussed, ideas about "hitch hiking" [6] and Haldane's "cost of selection" [11], and so on. Since the entire fossil history of digital organisms is preserved in the computer, it really should be possible to understand their evolution in quantitative terms.

But conversely, there are also potential benefits for evolutionary biology. In population genetics and evolutionary game theory we design models to study the success and failure of a prefined set of traits or strategies in the struggle for life. But what are the possible traits? And how well do they work out in *particular* environments with *particular* competitors? These questions are ignored in traditional models — they come in as parameters to be provided by developmental biology and ecology. For understanding the evolution of complex traits this is not satisfactory, because these parameters are themselves shaped by evolution. Evolutionary processes constantly shift the targets of evolutionary optimization, create spatial patterns, turn competitors into mutualists and create new levels of selection. Artificial Life models of such phenomena (e.g. [12, 13, 14]) promise to be useful for developing the concepts and techniques to deal with that challenge, but only if they are combined with the insights from almost a century of population genetics.

References

[1] R.E. Lenski, C. Ofria, R.T. Pennock, and C. Adami. The evolutionary origin of complex features. *Nature*, 8(423):139–144, May 2003.

[2] W.B. Provine. *Sewall Wright and evolutionary biology*. University of Chicago, Chicago, IL, 1986.

[3] T.S. Ray. An approach to the synthesis of life. In C.G. Langton, C. Taylor, J.D. Farmer, , and S. Rasmussen, editors, *Artificial Life II*, pages 371–408. Addison-Wesley, 1992.

[4] Karl Sims. Evolving virtual creatures. In *Computer Graphics (SIGGRAPH '94 Proceedings)*, pages 15–22. ACM Press, New York, NY, 1994.

[5] Hod Lipson and Jordan B. Pollack. Automatic design and manufacture of artificial lifeforms. *Nature*, 406:974–978, 2000.

[6] W.G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(269-294), 1966.

[7] J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genet.Res.*, 23:23–35, 1974.

[8] J. B. S. Haldane. *The causes of evolution*. Longmans, New York, 1932.

[9] Manfred Eigen. Self-organization of matter and the evolution of biological macro-molecules. *Naturwissenschaften*, 58:465–523, 1971.

[10] W. S. Stephan and D. A. Kirby. Rna folding in drosophila shows a distance effect for compensatory mutations. *Genetics*, 135:97–103, 1993.

[11] J.B.S. Haldane. The cost of natural selection. *Journal of Genetics*, 55:511–524, 1957.

[12] W. Daniel Hillis. Coevolving parasites improve simulated evolution as an optimization procedure. *Physica D*, 42:228–234, 1991.

[13] Maarten C. Boerlijst and Paulien Hogeweg. Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. *Physica D*, 48:17–28, 1991.

[14] R.A. Watson and J.B. Pollack. Symbiotic composition and evolvability. In Jozef Kelemen and Petr Sosik, editors, *Advances in Artificial Life*, volume 2159 of *Lecture Notes in Computer Science*, pages 480–490. Springer, Berlin, 2001.

Evolution, 57(6), 2003, pp. 1448–1450

COMMENTS

IS EVOLVABILITY INVOLVED IN THE ORIGIN OF MODULAR VARIATION?

ANDY GARDNER[1,2] AND WILLEM ZUIDEMA[1,3]

[1]*Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom*

[2]*E-mail: Andy.Gardner@ed.ac.uk*

[3]*Language Evolution and Computation Research Unit, Theoretical and Applied Linguistics, University of Edinburgh, 40 George Square, Edinburgh EH8 9LL, United Kingdom*

Abstract.—Lipson et al. (2002) presented an elegant linear algebraic formalism to define and study the evolution of modularity in an artificial evolving system. They employed simulation data to support their suggestion that modularity arises spontaneously in temporally fluctuating systems in response to selection for enhanced evolvability. We show analytically and by simulation that their correlate of modularity is itself under selection and so is not a reliable indicator of selection for modularity per se. In addition, we question the relation between modularity and evolvability in their simulations, suggesting that this modularity cannot confer enhanced evolvability.

Key words.—Adaptability, canalization, fluctuating selection, pleiotropy, robustness.

Received January 22, 2003. Accepted January 27, 2003.

Modularity is a major principle of design and abounds in nature. Functional separation of modules—from eukaryote organelles to *Drosophila* limbs to human cognitive faculties—may give robustness to changing inputs and facilitate future improvement. The question of the evolutionary origins of such modularity is important and the recent simulation study of Lipson et al. (2002) is therefore a welcome contribution. They introduce a potentially extremely useful formalism that allows one to quantify modularity and study its evolutionary origins. Environmental variables are described by a vector **E**, and phenotypic traits by a vector **P**. A matrix **A**, which premultiplies **E** to give **P**, then describes the organismal process of transforming environmental input into phenotypic output.

Lipson et al. argue that the "blockiness" of **A** and its correlate, the number of zero elements, are measures of modularity. By assigning fitnesses to realized phenotypes depending on their distance from an arbitrarily chosen optimum, Lipson et al. (2002) study the evolution of modularity. Their simulations show that the frequency of zero elements in the matrices deviates from the expected value (1/3, the frequency of zero elements at initialization and among random mutations) when the environment changes rapidly. Lipson et al. attribute these results to a "second order (delayed) pressure for decomposition for adaptability," (p.1554) that is, the uncoupling of traits to allow independent optimization of each and hence increased ability to adapt to new environments. Enhanced evolvability is concluded to be a cause, as well as a fortunate outcome, of the preponderance of zero-element-rich matrices. We disagree with this conclusion and believe that an alternative explanation exists. In addition, we feel that modularity cannot influence evolvability in their study.

In the simulations of Lipson et al., the element values of **E** are restricted to -1 and $+1$ and the element values of **A** are restricted to $-1, 0$, and $+1$. The elements of the phenotype vector **P** are therefore restricted to the range $-n \rightarrow n$, where n is the number of dimensions of the vectors (eight in the simulations of Lipson et al.). They restrict the elements of **F**, the arbitrary optimal phenotype, to -1 and $+1$. The optimal phenotypes are therefore restricted to a small subset of all possible phenotypes, centered on the origin. We find that matrices with many zero elements tend to produce phenotypes that are closer to the zero vector, and therefore on average closer to the optimal phenotypes (mathematical details are given in the Appendix).

Rather than appealing to enhanced evolvability, the preponderance of zero-rich matrices can be explained by the advantage delivered to any **A** that can maintain a phenotype close to the origin, despite environmental perturbation (i.e., canalization; Waddington 1942). In Figure 1 we give the probability distribution of the value of an element of **P** as a function of ζ, the number of zero elements in the corresponding row of **A**. As ζ increases, the value of the focal element of **P** is more tightly distributed about the origin. Figure 2 reveals the relation between ζ and the mean scalar residual (negatively correlated with Lipson et al.'s measure of fitness) in a focal dimension: increasing ζ reduces the residual and thus increases fitness. Conducting simulations of our own, we have been able to demonstrate frequencies of zero elements significantly greater than 1/3, even when mutation is suppressed. Hence, individual lineages may thrive or decline, but cannot evolve and therefore cannot be under selection for enhanced evolvability (see Fig. 3 and Table 1).

Moreover, in the set-up of Lipson et al., it is unclear why enhanced evolvability is expected to play any role. Each element of the vector **P** is the result of (dot-) multiplying a separate row vector from **A** with **E**. Contrary to the suggestions of Lipson et al., manipulating the elements of such a row vector has no effect on the value of other elements of **P**. This means that when evolving **A** in the context of a certain environment **E** and a certain target phenotype **F**, every element of the actual phenotype **P** can be optimized independently. Interestingly, a different use of the same formalism was suggested by Lipson et al. and avoids this problem. Under this alternative scheme, vector **E** describes the genotype and matrix **A** describes the genetic architecture of the phenotype (e.g., pleiotropy), a framework similar to the multiple quantitative trait model proposed by Taylor and Higgs (2000). By allowing both **E** and **A** to evolve, one can study the evolution of modularity and evolvability under, for example, fluctuations in **F**.

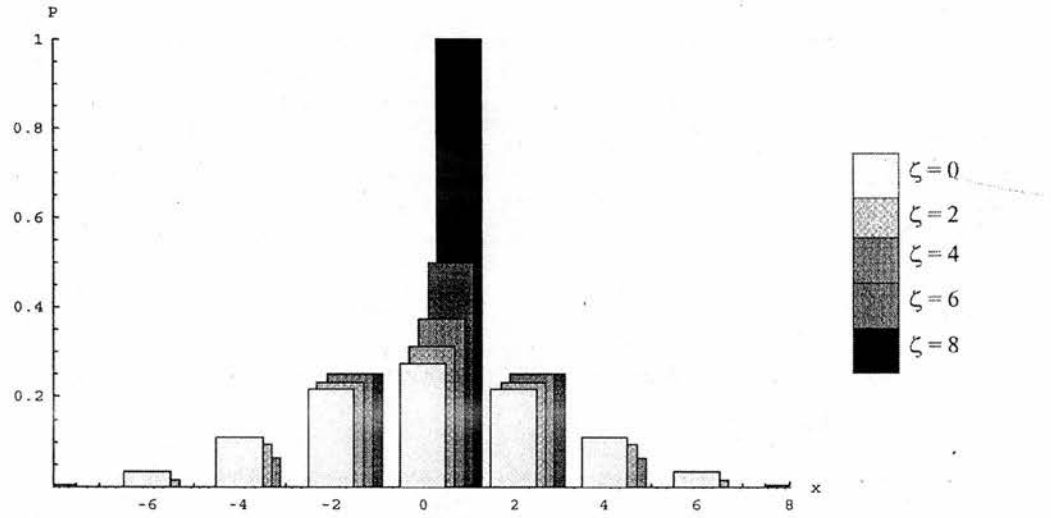


FIG. 1. The probability distribution of the value of $\mathbf{P}_\kappa$ as a function of the number of zero elements in the κ^{th} row of the 8 × 8 ternary matrix **A**, ζ. Here n (= 8) and every value of ζ (= 0, 2, 4, 6, 8) are even, so the values of $\mathbf{P}_\kappa$ are restricted to the set of even integers.

This is not to say that modularity is not under selection. It is possible that modularity confers robustness of fitness in response to the form of environmental change investigated by Lipson et al. When matrices are highly modular, such that there is a one-to-one correspondence between environmental characteristic and phenotypic trait, alteration of only one aspect of the environment will perturb the phenotype in one dimension only. Matrices that are less modular have environmental components each affecting more than one trait, and more than one trait being affected by several environmental components. They are therefore perturbed in multiple dimensions whenever a single aspect of the environment is altered. Because Lipson et al. change the sign of only one element of **E** at each environmental alteration, it is conceivable that selection for fitness robustness has given rise to an increase in modularity in their simulations. However, this is quite a different pressure than the supposed selection for enhanced evolvability.

In summary, Lipson et al. have presented an exciting and novel formalism that may yield quantitative, as well as qualitative insights into the evolution of evolvability and other problems. However, in their application of the model they have: (1) failed to demonstrate selection for modularity per se; and (2) not clearly established a link between modularity and evolvability. We suggest that enhanced evolvability can be neither a cause nor an outcome of the increase in their correlate of modularity.

ACKNOWLEDGMENTS

We thank N. Barton, A. Kalinka, and S. West for helpful discussion and comments on the manuscript, and H. Lipson for assistance in our re-creation of the evolutionary algorithm of the Lipson et al. (2002) paper. This work was supported

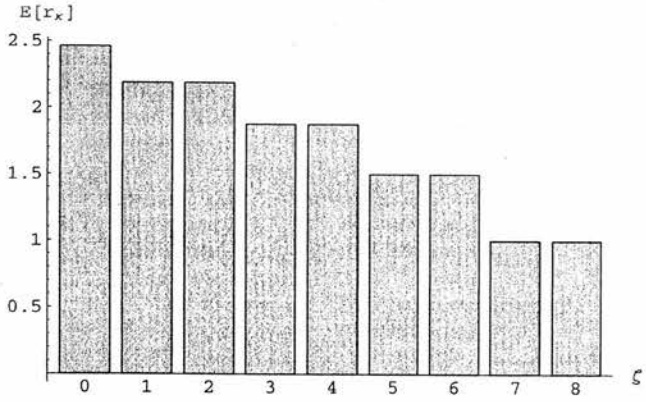


FIG. 2. The expectation of the residual r_κ as a function of ζ for an 8 × 8 ternary matrix. By ensuring that phenotype vectors are more tightly distributed around the origin, and hence closer to the optimum, matrix rows with more zero elements achieve reduced residual, on average.

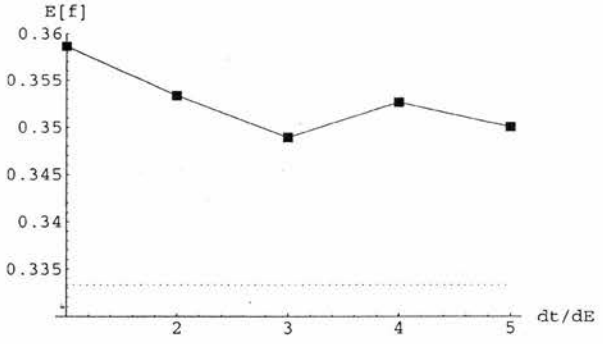


FIG. 3. The frequency of zero elements, averaged over 400 replicates, after 20 generations of evolution for a population of 50 8 × 8 matrices over a range of rates of environmental change dt/d**E**. The broken line indicates the null prediction 1/3. Simulations were devoid of mutation, but otherwise the evolutionary algorithm remained the same as that of Lipson et al.

TABLE 1. Simulation data and the one-tailed sign test for significant departure from null prediction "frequency of zero elements = 1/3".

dt/dE	Mean frequency of zero elements (from 400 replicates)	No. of replicates (out of 400) with frequency of zero elements >1/3	P
1	0.359	268	4.700×10^{-12}
2	0.353	243	9.979×10^{-6}
3	0.349	233	5.639×10^{-4}
4	0.353	250	3.266×10^{-7}
5	0.350	228	2.946×10^{-3}

by the Biotechnology and Biological Sciences Research Council (AG) and a Marie Curie fellowship from the European Commission (WZ).

LITERATURE CITED

Lipson, H., J. B. Pollack, and N. P. Suh. 2002. On the origin of modular variation. Evolution 56:1549–1556.

Taylor, C. F., and P. G. Higgs. 2000. A population genetics model for multiple quantitative traits exhibiting pleiotropy and epistasis. J. Theor. Biol. 203:419–437.

Waddington, C. H. 1942. Canalization of development and the inheritance of acquired characters. Nature 150:563–565.

Corresponding Editor: R. Harrison

APPENDIX

The Distribution of P_κ

A is a nxn ternary matrix (element values are -1, 0, and $+1$) and **E** is a n-element column vector with element values $+1$ and -1. The product of the premultiplication of **E** by **A** gives the phenotype vector **P**. The κ^{th} element of **P** is given by $\mathbf{P}_\kappa = \mathbf{A}_\kappa.\mathbf{E} = \Sigma_i \mathbf{A}_{\kappa i}.\mathbf{E}_i = \zeta.0 + m.(+1) + (n - \zeta - m).(-1)$ where ζ is the number of zero elements in $\mathbf{A}_\kappa$ and $m \sim Bin(n - \zeta, 1/2)$ is the number of same-sign pairs of $\mathbf{A}_{\kappa i}$ and $\mathbf{E}_i$ (i.e., those pairs of elements multiplying to give $+1$). Rearranging, the probability distribution of $\mathbf{P}_\kappa$ is found to be

$$P[\mathbf{P}_\kappa = x] = \binom{n - \zeta}{\frac{n - \zeta - x}{2}} 2^{\zeta - n}, \tag{A1}$$

for $n = 8$, the distribution of P_κ as a function of ζ is shown in Figure 1.

$E[r_\kappa]$ as a function of ζ

Lipson et al. define fitness as a decreasing function of the (scalar) distance between realized phenotype **P** and an arbitrary optimum **F**. The residual in the κ^{th} dimension is $r_\kappa = |\mathbf{F}_\kappa - \mathbf{P}_\kappa|$ where $\mathbf{F}_\kappa$ takes value $+1$ or -1 with equal probability. The probability density function of r_κ is then

$$P[r_\kappa = y] = \frac{1}{2}P[|\mathbf{P}_\kappa| - 1 = y] + \frac{1}{2}P[|\mathbf{P}_\kappa| + 1 = y]$$

$$= \frac{1}{2}\Big(P[|\mathbf{P}_\kappa| = y + 1] + P[|\mathbf{P}_\kappa| = y - 1]\Big). \tag{A2}$$

Because P_κ is symmetrical about the origin, $P[\mathbf{P}_\kappa = z] = P[\mathbf{P}_\kappa = -z]$ and so for $z > 0$, $P[|\mathbf{P}_\kappa| = z] = 2\,P[\mathbf{P}_\kappa = z]$, that is, for $y > 1$,

$$P[r_\kappa = y] = P[\mathbf{P}_\kappa = y + 1] + P[\mathbf{P}_\kappa = y - 1]. \tag{A3}$$

For $y \leq 1$;

$$P[r_\kappa = 1] = P[\mathbf{P}_\kappa = -2]P[F_\kappa = -1] + P[\mathbf{P}_\kappa = +2]P[F_\kappa = +1]$$
$$+ P[\mathbf{P}_\kappa = 0] = P[\mathbf{P}_\kappa = +2] + P[\mathbf{P}_\kappa = 0]$$

$$P[r_\kappa = 0] = P[\mathbf{P}_\kappa = -1]P[\mathbf{F}_\kappa = -1] + P[\mathbf{P}_\kappa = +1]P[\mathbf{F}_\kappa = +1]$$
$$= P[\mathbf{P}_\kappa = +1]. \tag{A4}$$

Because $r_\kappa = \mathbf{P}_\kappa \pm 1$, and $\mathbf{P}_\kappa$ is restricted to values of the same parity as $n - \zeta$, r_κ is only evaluated for those integers with parity opposite to $n - \zeta$. For $n = 8$, the mean of r_κ is revealed as a function of ζ in Figure 2.

How the poverty of the stimulus solves the poverty of the stimulus

Willem Zuidema
Theoretical and Applied Linguistics &
Institute for Cell, Animal and Population Biology
University of Edinburgh
40 George Square, Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

Abstract

Language acquisition is a special kind of learning problem because the outcome of learning of one generation is the input for the next. That makes it possible for languages to *adapt* to the particularities of the learner. In this paper, I show that this type of language change has important consequences for models of the evolution and acquisition of syntax.

1 The Language Acquisition Problem

For both artificial systems and non-human animals, learning the syntax of natural languages is a notoriously hard problem. All healthy human infants, in contrast, learn any of the approximately 6000 human languages rapidly, accurately and spontaneously. Any explanation of how they accomplish this difficult task must specify the (innate) *inductive bias* that human infants bring to bear, and the input data that is available to them. Traditionally, the inductive bias is termed – somewhat unfortunately – "Universal Grammar", and the input data "primary linguistic data".

Over the last 30 years or so, a view on the acquisition of the syntax of natural language has become popular that has put much emphasis on the innate machinery. In this view, that one can call the "Principles and Parameters" model, the Universal Grammar specifies most aspects of syntax in great detail [e.g. 1]. The role of experience is reduced to setting a limited number (30 or so) of parameters. The main argument for this view is the *argument from the poverty of the stimulus* [2]. This argument states that children have insufficient evidence in the primary linguistic data to induce the grammar of their native language.

Mark Gold [3] provides the most well-known formal basis to this argument. Gold introduced the criterion "identification in the limit" for evaluating the success of a learning algorithm: with an infinite number of training samples all hypotheses of the algorithm should be identical, and equivalent to the target. Gold showed that the class of context-free grammars is not learnable in this sense by any algorithm from positive samples alone (and neither are other *super-finite* classes). This proof is based on the fact that no matter how many samples from an infinite language a

learning algorithm has seen, the algorithm can not decide *with certainty* that the samples are drawn from the infinite language or from a finite language that contains all samples. Because natural languages are thought to be at least as complex as context-free grammars, and negative feedback is assumed to be absent in the primary linguistic data, Gold's analysis, and subsequent work in learnability theory [1], is usually interpreted as strong support for the argument from the poverty of the stimulus, and, in the extreme, for the view that grammar induction is fundamentally impossible (a claim that Gold would not subscribe to).

Critics of this "nativist" approach [e.g. 4, 5] have argued for different assumptions on the appropriate grammar formalism (e.g. stochastic context-free grammars), the available primary data (e.g. semantic information) or the appropriate learnability criterion. In this paper I will take a different approach. I will present a model that induces *context-free grammars* without a-priori restrictions on the search space, semantic information or negative evidence. Gold's negative results thus apply. Nevertheless, acquisition of grammar is successful in my model, because another process is taken into account as well: the cultural evolution of language.

2 The Language Evolution Problem

Whereas in language acquisition research the central question is how a child acquires an *existing* language, in language evolution research the central question is how this language and its properties have emerged in the first place. Within the nativist paradigm, some have suggested that the answer to this question is that Universal Grammar is the product of evolution under selection pressures for communication [e.g. 6]. Recently, several formal models have been presented to evaluate this view. For this paper, the most relevant of those is the model of Nowak et al. [7].

In that model it is assumed that there is a finite number of grammars, that newcomers (infants) learn their grammar from the population, that more successful grammars have a higher probability of being learned and that mistakes are made in learning. The system can thus be described in terms of the changes in the relative frequencies x_i of each grammar type i in the population. The first result that Nowak et al. obtain is a "coherence threshold". This threshold is the necessary condition for grammatical coherence in a population, i.e. for a majority of individuals to use the same grammar. They show that this coherence depends on the chances that a child has to correctly acquire its parents' grammar. This probability is described with the parameter q. Nowak et al. show analytically that there is a minimum value for q to keep coherence in the population. If q is lower than this value, all possible grammar types are equally frequent in the population and the communicative success in minimal. If q is higher than this value, one grammar type is dominant; the communicative success is much higher than before and reaches 100% if $q = 1$.

The second result relates this required fidelity (called q_1) to a lower bound (b_c) on the number of sample sentences that a child needs. Nowak et al. make the crucial assumption that all languages are equally expressive and equally different from each other. With that assumption they can show that b_c is proportional to the total number of possible grammars N. Of course, the actual number of sample sentences b is finite; Nowak et al. conclude that only if N is relatively small can a stable grammar emerge in a population. I.e. the population dynamics require a restrictive Universal Grammar.

The models of Gold and Nowak et al. have in common that they implicitly assume that every possible grammar is equally likely to become the target grammar for learning. If even the best possible learning algorithm cannot learn such a grammar,

the set of allowed grammars must be restricted. There is, however, reason to believe that this assumption is not the most useful for language learning. Language learning is a very particular type of learning problem, because the outcome of the learning process at one generation is the input for the next. The samples from which a child learns with its learning procedure, are therefore *biased* by the learning of previous generations that used the same procedure[8].

In [9] and other papers, Kirby, Hurford and students have developed a framework to study the consequences of that fact. In this framework, called the "Iterated Learning Model" (ILM), a population of individuals is modeled that can each produce and interpret sentences, and have a language acquisition procedure to learn grammar from each other. In the ILM one individual (the parent) presents a relatively small number of examples of form–meaning pairs to the next individual (the child). The child then uses these examples to induce his own grammar. In the next iteration the child becomes the parent, and a new individual becomes the child. This process is repeated many times. Interestingly, Kirby and Hurford have found that in these iterated transmission steps the language becomes easier and easier to learn, because the language adapts to the learning algorithm by becoming more and more structured. The structure of language in these models thus emerges from the iteration of learning. The role of biological evolution, in this view, is to shape the learning algorithms, such that the complex results of the iterated learning is biologically adaptive [10]. In this paper I will show that if one adopts this view on the interactions between learning, cultural evolution and biological evolution, the models such as those of Gold [3] and Nowak et al. [7] can no longer be taken as evidence for an extensive, innate pre-specification of human language.

3 A Simple Model of Grammar Induction

To study the interactions between language adaptation and language acquisition, I have first designed a grammar induction algorithm that is simple, but can nevertheless deal with some non-trivial induction problems. The model uses context-free grammars to represent linguistic abilities. In particular, the representation is limited to grammars G where all rules are of one of the following forms: (1) $A \mapsto t$, (2) $A \mapsto BC$, (3) $A \mapsto Bt$. The nonterminals A, B, C are elements of the non-terminal alphabet V_{nt}, which includes the start symbol S. t is a string of terminal symbols from the terminal alphabet V_t[1]. For determining the language L of a certain grammar G I use simple depth-first exhaustive search of the derivation tree. For computational reasons, the depth of the search is limited to a certain depth d, and the string length is limited to length l. The set of sentences ($L' \subseteq L$) used in training and in communication is therefore finite (and strictly speaking not context-free, but regular); in production, strings are drawn from a uniform distribution over L'.

The grammar induction algorithm learns from a set of sample strings (sentences) that are provided by a teacher. The design of the learning algorithm is originally inspired by [11] and is similar to the algorithm in [12]. The algorithm fits within a tradition of algorithms that search for compact descriptions of the input data [e.g. 13, 14, 15]. It consists of three operations:

Incorporation: *extend the language, such that it includes the encountered string*; if string s is not already part of the language, add a rule $S \mapsto s$ to the grammar.

[1] Note that the restrictions on the rule-types above do not limit the scope of languages that can be represented (they are essentially equivalent to Chomsky Normal Form). They are, however, relevant for the language acquisition algorithm.

Compression: *substitute frequent and long substrings with a nonterminal, such that the grammar becomes smaller and the language remains unchanged*; for every valid substring z of the right-hand sides of all rules, calculate the compression effect $v(z)$ of substituting z with a nonterminal A; replace all valid occurrences of the substring $z' = argmax_z v(z)$ with A if $v(z') > 0$, and add a rule $A \mapsto z'$ to the grammar. "Valid substrings" are those substrings which can be replaced while keeping all rules of the forms 1–3 described above. The compression effect is measured as the difference between the number of symbols in the grammar before and after the substitution. The compression step is repeated until the grammar does not change anymore.

Generalization: *equate two nonterminals, such that the grammar becomes smaller and the language larger*; for every combination of two nonterminals A and B ($B \neq S$), calculate the compression effect v of equating A and B. Equate the combination $(A', B') = argmax_{A,B} v(A, B)$ if $v(A', B') > 0$; i.e. replace all occurrences of B with A. The compression effect is measured as the difference between the number of symbols before and after replacing and deleting redundant rules. The generalization step is repeated until the grammar does not change anymore.

4 Learnable and Unlearnable Classes

The algorithm described above is implemented in C^{++} and tested on a variety of target grammars[2]. I will not present a detailed analysis of the learning behavior here, but limit myself to a simple example that shows that the algorithm can learn some (recursive) grammars, while it can not learn others. The induction algorithm receives three sentences (abcd, abcabcd, abcabcabcd). The incorporation, compression (repeated twice) and generalization steps yield subsequently the following grammars:

(a) Incorporation	(b) Compression	(c) Generalization
S ↦ abcd	S ↦ Yd	S ↦ Xd
S ↦ abcabcd	S ↦ Xd	S ↦ Xabcd
S ↦ abcabcabcd	S ↦ Xabcd	X ↦ XX
	X ↦ YY	X ↦ abc
	Y ↦ abc	

In (b) the substrings "abcabc" and "abc" are subsequently replaced by the non-terminals X and Y. In (c) the non-terminals X and Y are equated, which leads to the deletion of the second rule in (b). One can check that the total size of the grammar reduces from 24, to 19 and further down to 16 characters.

From this example it is also clear that learning is not always successful. Any of the three grammars above ((a) and (b) are equivalent) could have generated the training data, but with these three input strings the algorithm always yields grammar (c). Consistent with Gold's general proof [3], many target grammars will never be learned correctly, no matter how many input strings are generated. In practice, each finite set of randomly generated strings from some target grammar, might yield a different result. Thus, for some number of input strings T, some set of target grammars are always acquired, some are never acquired, and some are some of the time acquired. If we can enumerate all possible grammars, we can describe this with a matrix Q, where each entry Q_{ij} describes the probability that the algorithm learning from sample strings from a target grammar i, will end up with grammar

[2]The source code is available at `http://www.ling.ed.ac.uk/~jelle`

of type j. Q_{ii} is the probability that the algorithm finds the target grammar. To make learning successful, the target grammars that are presented to the algorithm have to be biased. The following section will show that for this we need nothing more than to assume that the output of one learner is the input for the next.

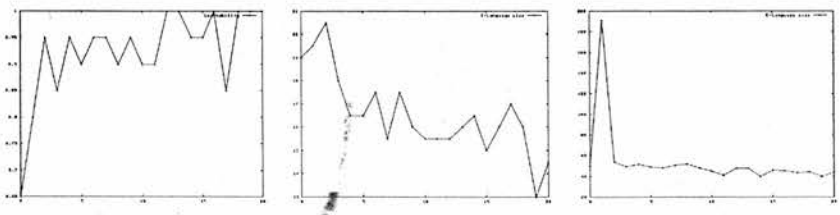
5 Iterated Learning: the Emergence of Learnability

To study the effects of iterated learning, we extend the model with a population structure. In the new version of the model individuals (agents, that each represent a generation) are placed in a *chain*. The first agent induces its grammar from a number E of randomly generated strings. Every subsequent agent (the child) learns its grammar from T sample sentences that are generated by the previous one (the parent). To avoid insufficient expressiveness, we also extend the generalization step with a check if the number E_G of different strings the grammar G can recognize is larger than or equal to E. If not, $E - E_G$ random new strings are generated and incorporated in the grammar. Using the matrix Q from the previous section, we can formalize this *iterated learning model* with the following general equation, where x_i is the probability that grammar i is the grammar of the current generation:

$$\Delta x_i = \sum_{j=0}^{N} x_j Q_{ji} \tag{1}$$

In simulations such as the one of figure 1 communicative success between child and parent – a measure for the learnability of a grammar – rises steadily from a low value (here 0.65) to a high value (here 1.0). In the initial stage the grammar shows no structure, and consequently almost every string that the grammar produces is idiosyncratic. A child in this stage typically hears strings like "ada", "ddac", "adba", "bcbd", or "cdca" from its parent. It can not discover many regularities in these strings. The child therefore can not do much better than simply reproduce the strings it heard (i.e. T random draws from at least E different strings), and generate random new strings, if necessary to make sure its language obeys the minimum number (E) of strings. However, in these randomly generated strings, sometimes regularities appear. I.e., a parent may use the randomly generated strings "dcac", "bcac", "caac" and "daac". When this happens the child tends to analyze these strings as different combinations with the building block "ac". Thus, typically, the learning algorithm generates a grammar with the rules $S \mapsto dcX$, $S \mapsto bcX$, $S \mapsto caX$, $S \mapsto daX$, and $X \mapsto ac$. When this happens to another set of strings as well, say with a new rule $Y \mapsto b$, the generalization procedure can decide to equate the non-terminals X and Y. The resulting grammar can then generalize from the observed strings, to the unobserved strings "dcb", "bcb", "cab" and "dab". The child still needs to generate random new strings to reach the minimum E, but fewer than in the case considered above.

The interesting aspect of this becomes clear when we consider the next step in the simulation, when the child becomes itself the parent of a new child. This child is now presented with a language with more regularities than before, and has a fair chance of *correctly* generalizing to unseen examples. If, for instance, it only sees the strings "dcac", "bcac", "caac", "bcb", "cab" and "dab", it can, through the same procedure as above, infer that "daac" and "dcb" are also part of the target language. This means that (i) the child shares more strings with its parent than just the ones it observes and consequently shows a higher between generation communicative success, and (ii) regularities that appear in the language by chance, have a fair chance to remain in the language. In the process of iterated learning, languages can thus become more structured and better learnable.



(a) Learnability (b) Number of rules (c) Expressiveness

Figure 1: Iterated Learning: although initially the target language is unstructured and difficult to learn, over the course of 20 generations (a) the learnability (the fraction of successful communications with the parent) steadily increases, (b) the number of rules steadily decreases (combinatorial and recursive stategies are used), and (c) after a initial phase of overgeneralization, the expressiveness remains close to its minimally required level. Parameters: $V_t = \{a, b, c, d\}$, $V_{nt} = \{S, X, Y, Z, A, B, C\}$, T=30, E=20, l_0=3. Shown are the average values of 2 simulations.

Similar results with different formalisms were already reported before [e.g. 11, 16], but here I have used context-free grammars and the results are therefore directly relevant for the interpretation of Gold's proof [3]. Whereas in the usual interpretation of that proof [e.g. 1] it is assumed that we need innate constraints on the *search space* in addition to a smart *learning procedure*, here I show that even a simple learning procedure can lead to successful acquisition, because restrictions on the search space automatically emerge in the iteration of learning. If one considers learnability a *binary* feature – as is common in generative linguistics – this is a rather trivial phenomenon: languages that are not learnable will not occur in the next generation. However, if there are gradations in learnability, the cultural evolution of language can be an intricate process where languages get shaped over many generations.

6 Language Adaptation and the Coherence Threshold

When we study this effect in a version of the model where *selection* does play a role, it is also relevant for the analysis in [7]. The model is therefore extended such that at every generation there is a population of agents, agents of one generation communicate with each other and the expected number of offspring of an agent (the *fitness*) is determined by the number of successful interactions it had. Children still acquire their grammar from sample strings produced by their parent. Adapting equation 1, this system can now be described with the following equation, where x_i is now the relative fraction of grammar i in the population (assuming an infinite population size):

$$\Delta x_i = \sum_{j=0}^{N} x_j f_j Q_{ji} - \phi x_i \tag{2}$$

Here, f_i is the *relative fitness* (quality) of grammars of type i and equals $f_i = \sum_j x_j F_{ij}$, where F_{ij} is the expected communicative success from an interaction between an individual of type i and an individual of type j. The relative fitness f of a grammar thus depends on the frequencies of all grammar types, hence it is *frequency*

dependent. ϕ is the average fitness in the population and equals $\phi = \sum_i x_i f_i$. This term is needed to keep the sum of all fractions at 1. This equation is essentially the model of Nowak et al. [7]. Recall that the main result of that paper is a "coherence threshold": a minimum value for the learning accuracy q to keep coherence in the population. In previous work [unpublished] I have reproduced this result and shown that it is robust against variations in the Q-matrix, as long as the value of q (i.e. the diagonal values) remains equal for all grammars.

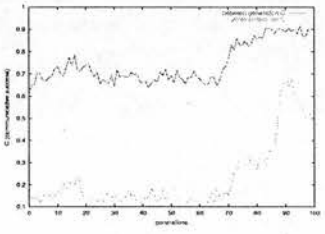


Figure 2: Results from a run under fitness proportional selection. This figure shows that there are regions of grammar space where the dynamics are apparently under the "coherence threshold" [7], while there are other regions where the dynamics are above this threshold. The parameters, including the number of sample sentences T, are still the same, but the language has adapted itself to the **bias** of the learning algorithm. Parameters are: $V_t = \{0, 1, 2, 3\}$, $V_{nt} = \{S, a, b, c, d, e, f\}$, P=20, T=100, E=100, l_0=12. Shown are the average values of 20 agents.

Figure 2, however, shows results from a simulation with the grammar induction algorithm described above, where this condition is violated. Whereas in the simulations of figure 1 the target languages have been relatively easy (the initial string length is short, i.e. 6), here the learning problem is very difficult (initial string length is long, i.e. 12). For a long period the learning is therefore not very successful, but around generation 70 the success suddenly rises. With always the same T (number of sample sentences), and with always the same grammar space, there are regions where the dynamics are apparently under the "coherence threshold", while there are other regions where the dynamics are above this threshold. The language has adapted to the learning algorithm, and, consequently, the coherence in the population does not satisfy the prediction of Nowak et al.

7 Conclusions

I believe that these results have some important consequences for our thinking about language acquisition. In particular, they offer a different perspective on the argument from the poverty of the stimulus, and thus on one of the most central "problems" of language acquisition research: *the logical problem of language acquisition.* My results indicate that in *iterated learning* it is not necessary to put the (whole) explanatory burden on the representation bias. Although the details of the grammatical formalism (context-free grammars) and the population structure are deliberately close to [3] and [7] respectively, I do observe successful acquisition of grammars from a class that is unlearnable by Gold's criterion. Further, I observe grammatical coherence even though many more grammars are allowed in principle than Nowak et al. calculate as an upper bound. The reason for these surprising results is that language acquisition is a very particular type of learning problem: it is a problem where the target of the learning process is itself the outcome of a learning process. That opens up the possibility of language itself to adapt to the

language acquisition procedure of children. In such iterated learning situations [11], learners are only presented with targets that other learners have been able to learn.

Isn't this the traditional Universal Grammar in disguise? Learnability is – consistent with the undisputed proof of [3] – still achieved by constraining the set of targets. However, unlike in usual *interpretations* of this proof, these constraints are not strict (some grammars are better learnable than others, allowing for an infinite "Grammar Universe"), and they are not a-priori: they are the outcome of iterated learning. The poverty of the stimulus is now no longer a problem; instead, the ancestors' poverty is the solution for the child's.

Acknowledgments This work was performed while I was at the AI Laboratory of the Vrije Universiteit Brussel. It builds on previous work that was done in close collaboration with Paulien Hogeweg of Utrecht University. I thank her and Simon Kirby, John Batali, Aukje Zuidema and my colleagues at the AI Lab and the LEC for valuable hints, questions and remarks. Funding from the Concerted Research Action fund of the Flemish Government and the VUB, from the Prins Bernhard Cultuurfonds and from a Marie Curie Fellowship of the European Commission are gratefully acknowledged.

References

[1] Stefano Bertolo, editor. *Language Acquisition and Learnability.* Cambridge University Press, 2001.
[2] Noam Chomsky. *Aspects of the theory of syntax.* MIT Press, Cambridge, MA, 1965.
[3] E. M. Gold. Language identification in the limit. *Information and Control (now Information and Computation)*, 10:447–474, 1967.
[4] Michael A. Arbib and Jane C. Hill. Language acquisition: Schemas replace universal grammar. In John A. Hawkins, editor, *Explaining Language Universals.* Basil Blackwell, New York, USA, 1988.
[5] J. Elman, E. Bates, et al. *Rethinking innateness.* MIT Press, 1996.
[6] Steven Pinker and Paul Bloom. Natural language and natural selection. *Behavioral and brain sciences*, 13:707–784, 1990.
[7] Martin A. Nowak, Natalia Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291:114–118, 2001.
[8] Terrence Deacon. *Symbolic species, the co-evolution of language and the human brain.* The Penguin Press, 1997.
[9] S. Kirby and J. Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer Verlag, London, 2002.
[10] Kenny Smith. Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior*, 2003. to appear.
[11] Simon Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight et al., editors, *The Evolutionary Emergence of Language.* Cambridge University Press, 2000.
[12] J. Gerard Wolff. Language acquisition, data compression and generalization. *Language & Communication*, 2(1):57–89, 1982.
[13] A. Stolcke. *Bayesian Learning of Probabilistic Language Models.* PhD thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley, 1994.
[14] Menno van Zaanen and Pieter Adriaans. Comparing two unsupervised grammar induction systems: Alignment-based learning vs. EMILE. In Ben Kröse et al., editors, *Proceedings of BNAIC 2001*, 2001.
[15] Zach Solan, Eytan Ruppin, David Horn, and Shimon Edelman. Automatic acquisition and efficient representation of syntactic structures. *This volume.*
[16] Henry Brighton. Compositional syntax from cultural transmission. *Artificial Life*, 8(1), 2002.

Modeling language acquisition, change and variation

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute of Animal, Cell and Population Biology
University of Edinburgh
40, George Square
Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

http://www.ling.ac.uk/ ~ jelle

Abstract

The relation between Language Acquisition, Language Change and Language Typology is a fascinating topic, but also one that is difficult to model. I focus in this paper on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and language change is parameter change. I review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach that is based on "Explicit Induction" algorithms for grammatical formalisms. I discuss which approach is most useful for which problems.

1 Language acquisition, change and typology

Every healthy human infant is capable of acquiring any one of a dazzling variety of human languages. This simple fact poses two fundamental challenges for linguistics: (1) understanding how children are so extremely successful at this apparently complex task, and (2) understanding how, although all humans have such similar linguistic abilities, such a wide variety of languages has emerged. These challenges are intricately linked: the languages that we observe today, are the result of thousands of years of cultural transmission, where every generation has acquired its language from the observed use by previous generations. That makes the acquisition of language a rather unique learning problem for learning theory, because what is being learned is itself the result of a learning process. Conversely, the structure of a language (say modern English) at any one time (say, 2003) is the result of perhaps millions of individuals learning from examples from a language with a very similar structure (say, the English of the 1960s).

This so-called *circular causality* (Steels, 1999) makes the relation between Language Acquisition, Language Change and Language Typology a fascinating topic, but also one that is difficult to model. I will focus here on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and

language change is parameter change. In the following I will review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach from the emerging field of computational modeling of the evolution of language (Kirby, 2002b), that is based on "Explicit Induction" algorithms for grammatical formalisms. I will argue that the differences between the two approaches have been exaggerated, and will discuss for which sort of problems which sort of approach is most useful.

2 Parameter models

The "Parameter change" approach to this problem is based on *parameterizing* linguistic structure, such that we can characterize all differences between possible human languages by a vector of a small number of parameters. E.g., in the Principles and Parameters approach (Chomsky, 1981; Bertolo, 2001), language acquisition is described in terms of parameter settings for a universal core, the Universal Grammar. With such a description of language in hand, we can reformulate the challenges as follows: (1) how can learning, given primary linguistic data that conforms to any particular set of parameters, find that set of parameters? (2) given a set of learning procedures that are capable of finding the correct parameters, which ones predict the type of language change and statistical distributions (universals tendencies, Kirby 1999) that we can actually observe?

2.1 Parameter setting

In the "parameter setting" models of language acquisition, one assumes a finite number N of possible grammars. If all variation can be described by n different, Boolean and independent parameters, such that the total number of possible grammars is $N = 2^n$. Such parameters determine, for instance, whether or not an object precedes the main verb in a sentence, or whether or not the subject can be left out. Typically, although the number of parameters is estimated at around 30, concrete examples are only worked out for the 2 or 3 least controversial proposed parameters. A lot of work in parameter setting works with rather simplified models that can be studied analytically, and that depend only on the finiteness of N. Examples of such models are "memory-less learning", "batch learning" (e.g. Nowak *et al.*, 2001) and "learning by enumeration" (Gold, 1967). It is useful to look in a bit more detail at these models.

Memory-less learning (Niyogi, 1998) is arguably the simplest language acquisition model. The algorithm works by choosing a random grammar from the set of possible grammars each time the input data shows that the present hypothesis is wrong. The algorithm obviously is not very efficient, because it can arrive at hypotheses it has already rejected before; i.e. each time it randomly chooses a new grammar, it forgets what it has learned from all data it has received before. This algorithm is only of interest because it is simple and provides a lower bound on the performance of any reasonable algorithm (Nowak *et al.*, 2001).

The *batch learner*, in contrast, memorizes all received sentences and finds all grammars from the set of possible ones that are consistent with these sentences. Equivalently, it keeps track of all possible grammars that are still consistent with the received data. In any case, for any reasonably large set of possible grammars, the batch learner has monstrous memory and processing requirements. Its value lies in the fact that it is simple, and provides an upper bound on the performance of any reasonable learning algorithm, as long as there is no a-priori reason to prefer one grammar that is consistent with the data over another.

As exemplified by appendix A, we can, with a bit of effort, derive explicit formulas that describe the probability of success q as a function of the number of input sentences for both the memory-less and the batch learner. Under the assumption that every wrong grammar is equally similar to the right grammar (described with a similarity parameter a), we can in fact give a complete transition matrix T, where all

diagonal values are $q_{memoryless}$ and all off-diagonal values are $(1 - q_{memoryless})/(N - 1)$. This transition matrix plays an important role in models of language change described in the next section.

It is important to realize that these algorithms only work because a finite (and in fact, relatively small) number of possible grammars is assumed. Moreover, calculations such as in appendix A are relatively easy due to some important assumptions: (1) that the algorithms are not biased at all to favor certain possible grammars over others; (2) that (in the case of the memory-less learner) the probability of jumping to a wrong or right grammar remains constant throughout the learning process; and (3) that all grammars are equally similar to each other. Without these assumptions, similar calculations quickly get rather complex.

For instance, *learning by enumeration* (Gold, 1967), as the name suggests, proceeds by enumerating one at a time, and in prespecified order all possible grammars. Only if a grammar is inconsistent with incoming data ("text"), does the algorithm move on to the next grammar. The procedure is of interest, because it can be used as a criterion for learnability (Gold, 1967)[1]. Calculating q is more difficult than before, because the probably of changing to a wrong grammar *decreases* over time.

The *trigger learning algorithm* (Wexler & Culicover, 1980) is a popular model that is of (slightly) more practical interest. Rather than picking a random new grammar, as the memory-less learner does, or enumerating grammars in a random order, as in learning by enumeration, it changes a random parameter when it finds an input sentence that is inconsistent with the present hypothesis. If with the new parameter setting the sentence can be parsed, the change is kept, otherwise it is reverted. The trigger learning algorithm thus implements a kind of hill-climbing (gradient ascent), by keeping parameters that do well and only making a small change when it improves performance. The probability of the trigger algorithm to give the right grammar after b sentences is even more tricky to calculate, because the probability to reject a wrong hypotheses *decreases* as more and more parameters get correctly set.

Many other parameter setting models exists. E.g. Briscoe (2002a) develops a variant of the trigger learning algorithm, where parameters are no longer independent, but fall into linguistically motivated inheritance hierarchies. Further, rather than choosing a single parameter at random and changing it, as in the TLA, Briscoe's algorithm selects several random parameters and keeps track of their most likely setting in a Bayesian, statistical fashion. Yang (2000) argues that language acquisition is best viewed as a selectionist process, where many different parameter sets are considered in parallel. Niyogi & Berwick (1995) and Yang (2000) consider the further complication that children learn from input sentences that are drawn from different languages, and explore the expectations on what grammar settings they will end up with. In all these models, calculating the probabilities of the outcome of learning gets very complex and results are typically obtained by using computer simulations.

2.2 Parameter change

Niyogi & Berwick (1995), as well as neural network modelers Hare & Elman (1995), argue that a theory of language acquisition – and the mistakes children make when confronted with insufficient or ambiguous input – implies a theory of language change. Similarly, Kirby (1999) explores the idea that a theory on language use and processing – which alter the primary linguistic data – leads to specific expectations on language change and the resulting linguistic variation. Hence, by working out the consequences for language change and comparing them to empirical data, theories on language use, processing and acquisition can be

[1]Learning by enumeration can, within finite time, find the target grammar from a class of grammars if the following conditions hold: (1) the class of grammars is finite (enumerable), (2) for every two grammars in the class, there exists a sentence that distinguishes between two grammars (i.e. that is grammatical according to one, and ungrammatical according to the other), and (3) the distinguishing sentence will occur within a finite amount of time in the the text, generated by the target grammar. It follows that the class of grammars is then learnable from text. It can be shown that superfinite classes of grammars, such as the context-free or context-sensitive grammars, are not learnable in this sense (Gold, 1967). Principles & Parameters-models, in contrast, are learnable (Wexler & Culicover, 1980) and so are many other classes (Angluin, 1980).

tested. Formally, a class of grammars $\mathcal{G}$, a learning algorithm $\mathcal{A}$ and a model of the primary linguistic data (a probability distribution $\mathcal{P}_i$ over the possible sentences of language i) together constitute the main ingredients of a dynamical system that describes the change in numbers of speakers of each language[2].

Several general results have been obtained. For instance, Niyogi & Berwick (1995) and Yang (2000) find that with different choices for $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$, the change in the number of speakers of a particular language tends to follow an S-shaped curve, consistent with observed patterns in historical data. More interestingly, Nowak *et al.* (2001) derive a *coherence threshold*. In their model, natural selection selecting for more frequent grammars, helps a population to converge on a specific grammar. Mistakes in learning, on the other hand, lead to divergence, because it essentially randomizes the choice of grammars. Nowak et al. find that if the accuracy in learning is below a precise threshold, all coherence in the population is lost and all languages are spoken with equal probability[3].

Niyogi and Berwick apply their methodology to a number of case studies. For instance, they look at a simple 3-parameter system where the parameters determine whether or not specifiers (1) and complements (2) come before the head of a phrase, and whether or not the verb is obligatorily in second position (3). In this system, there are 8 different possible grammars (languages). By making assumptions on the frequency with which triggers for each of the parameters are available to the child, they can estimate the probability a specific learning algorithm can learn each language. They numerically determine the probabilities of transitions between each of the 8 language over 30 generations with 128 triggers per generation. They find that languages with the third parameter set to "0" ($V2-$) are extremely unstable and that the $V2+$ parameter therefore quickly gets fixed in all simulations. This observation is contrary to observed trends in historical data, where $V2+$ is typically lost. Niyogi and Berwick argue that this falsifies their preliminary model, and thus illustrates the feasibility of testing the diachronic accuracy of the assumptions on $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$.

2.3 Some features of parameter change models

Several other parameter change models have been studied. They have in common the emphasis on the uniformity of languages, i.e. all possible languages (grammars) are of equal quality. Hence, children acquiring a language do not go from a simple grammar to a more complex one, but rather jump from one grammar to an equally complex alternative. Not the quality of the language, but the uncertainty about which is the correct one changes over time.

Moreover, in all these models the acquisition of syntax is studied independently from the acquisition of phonology, semantics, pragmatics and the lexicon, and, usually, independent from the particularities of the child's parsing algorithm. The training data are "triggers", i.e. strings of grammatical categories. The problems of learning the syntactic categories of words and their meaning, and learning to recognize the phonological form and the boundaries between words are all ignored.

Further, the models fit into a tradition that is much mathematically oriented. Although many results are obtained through numerical simulations, the models are formulated at a rather abstract level. Generations are typically discrete, the number of parameters small (2, 3, 5), number of training samples and the number of individuals in a population very small or, alternatively, infinite.

The models are valuable, because they give a *general* insight in how linguistic conventions can change and spread in a population. However, the problem with this approach is that its potential for explaining *specific* aspects of language acquisition and language typology depends completely on the successful parametrization of linguistic descriptions. That dependence has advantages, because it makes the relation with other linguistic theories very clear, but it has some major disadvantages as well.

[2]In addition to the triple $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$ (Niyogi & Berwick, 1995), one needs assumptions on population and generation structure and the number of training sentences the algorithm receives.

[3]Presumably, a similar mechanism explains the lack of coherence in the simulations of Niyogi & Berwick (1995).

First, there is, as for now, no such parametrization available. If efficient parametrization (i.e. with 20 or 30 parameters) turns out to be impossible, models that depend on them will be inadequate. Second, even if it is possible in principle, without a complete theory available on what each parameter means, solutions in terms of these parameters give little insight on why children learn certain things with more ease than others, or why languages tend to show certain patterns more often than others. Finally, parameter-models might give an adequate description of the variation in languages in a quasi-stable state, but that does not necessarily mean that they also give an adequate description of language variety when languages are changing. In particular, observed trends in language change regarding the interaction between phonology, syntax, semantics and pragmatics seem hard to capture in available parameter models.

3 Explicit Induction

3.1 Grammar Induction: impossible and irrelevant?

Grammar Induction algorithms are usually based on the intuition that the frequency of occurrence of substring in the training sentences, and the contexts in which they appear, contain information on what the underlying constituents and the rules of combination of the target grammar are. E.g. Zellig Harris, in describing the methods linguists use to infer the grammar of an unknown language, defines the crucial concept of "substitutability" as follows: "If our informant accepts DA'F as a repetition of DEF, and if we are similarly able to obtain E'BC as equivalent to ABC, then we say that A and E are mutually substitutable" (Zellig Harris, 1951, quoted in van Zaanen 2001).

It is possible to design induction algorithms that, just like Harris's linguist, use observed patterns in training sentences to induce the underlying grammar. However, due to initial negative results on the theoretical possibility of learning a grammar from positive data (Gold, 1967) and developments in linguistic theory (e.g. Chomsky, 1965), the *induction* of grammar has been widely viewed as both impossible and irrelevant.

The supposed impossibility of grammar induction is based on a widespread misinterpretation of negative learnability results. Gold (1967) showed that e.g. the class of context-sensitive languages is not *identifiable in the limit*. Even we if accept identification in the limit as the appropriate criterion for learnability, Gold's results mean nothing more than, in his own words:

> *"The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally. Equivalently, we may say that the child starts out with more information than that the language it will be presented is context-sensitive. In particular, the results on learnability from text imply the following: The class of possible natural languages if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality." (Gold, 1967)*

In other words, a class of context-sensitive grammars needs to be constrained to make it learnable. Angluin (1980) has shown that very non-trivial classes of formal languages are learnable. Nothing in the formal results, however, proves that the necessary restrictions are due to an extensive, innate, language-specific Universal Grammar; they could be simply generic properties of the human brain[4].

The supposed irrelevance of grammar induction algorithms is based on the fact that the dominant linguistic theories of the last decades assume extensive innate knowledge. If children don't do grammar induction, why design computer programs that do? Evidence for this view comes – in addition to the learnability

[4] Although it is of course true that learnability is a valid test for judging the validity of a (grammatical) theory, and that few proposed theories other than those from the nativist tradition pass it. However, one can argue that nativist theories, rather than solving the learnability problem, simply shift it to the domains of evolutionary theory and cognitive neuroscience.

results – from empirical observations in child language acquisition. Typically, such arguments have the form: the child correctly uses construction X very early in life, even though the primary linguistic data it has received up to that point does not provide enough evidence to choose between X and several alternative logical possibilities. Thus, it is concluded, the child must have prior (innate) knowledge of X.

More and more it is now recognized that this "knowledge of X" might be an emergent result of the interaction between not necessarily language-specific cognitive and learning abilities, and the structure, meaning and pragmatics of the linguistic data the child received (MacWhinney, 1999). Consequently, the need to postulate language-specific adaptations might be limited (Jackendoff, 2002; Hauser *et al.*, 2002).

3.2 Induction Algorithms

Wolff (1982), and similarly Stolcke (1994), Langley & Stromsten (2000) and Zuidema (2003), presents a model based on the idea that a grammar is a compressed representation of a possibly infinite language (string set). These algorithms all use context-free grammars as the grammar formalism, learn from text and run through three phases that can be termed "incorporation", "compression" and "generalization". I will refer to these algorithms as "compression-based induction".

In the incorporation phase, input sentences s are stored as idiosyncratic rewrite rules $S \mapsto s$. In the compression phase (or "syntagmatic merging"), the most frequent substrings z in the right-hand sides of the stored rules are replaced by a unique non-terminal symbol N. Rules of the form $N \mapsto z$ are added to the grammar. In the generalization phase (or "paradigmatic merging"), two nonterminals N and N' are considered *substitutable* if they occur in the same context; all occurrences of N' are then replaced by N. Different variants of the basic algorithm differ in how *greedy* they are, and in whether or not they are *incremental.* Kirby (2000), and later papers, uses a algorithm were the context-free grammars are enriched with a predicate-logic based semantics.

A related framework based on substitutability is developed by van Zaanen (2001) and termed "Alignment Based Learning" (ABL). Van Zaanen develops a number of algorithms for the two phases of the ABL framework: Alignment learning and selection learning. In the alignment learning phase input sentences are compared, aligned and common substrings are identified. The *unequal* parts z and z' of the two sentences are labeled with a non-terminal. The non-terminal is unique if neither z nor z' was labeled already, but the algorithm reuses the existing label if available, and equates the two non-terminals if both z and z' were labeled already. In the latter two conditions a form of generalization occurs. Each labeling is a hypothesis on a possible constituent of the target language, and very many such hypotheses are generated.

In the selection learning phase, a subset of the generated hypotheses is selected. That subset is chosen such that it is concise (each hypothesis can be used to analyze many sentences), and that it is internally consistent (hypotheses do not overlap). The ABL algorithm yields a tree-bank: an annotated version of the input corpus (it thus implements automated tagging). From the tree-bank, context-free grammars can be trivially induced.

3.3 Language Evolution

In the "Explicit Induction" approach to modeling language change and evolution, language change is studied based on similar induction algorithms, i.e. learning algorithms that produce an explicit grammar based on training sentences (see Hurford, 2002, for a review). Such an approach avoids the problems of parameter models, because they can incorporate any available linguistic formalism. However, they have two major disadvantages as well: (1) language induction is very challenging problem that is far from solved, even for simplified and well understood grammar formalisms; (2) models that incorporate a full-blown linguistic formalism, including procedures for language production and interpretation, quickly get very complex.

Two recent models by Kirby (2002a) and Batali (2002) show that there is reason for optimism for progress on bl problems. Kirby presents a model that is very clear in its set-up. It uses first-order predicate logic with a small set of entities and predicates to represent semantics, and a extension of context-free grammars to represent syntax and the syntax-semantics mapping. The model thus uses well-understood and conventional linguistic formalisms and a simple learning procedure. However, by using the output of one learning cycle as input for the next Kirby was able to get some unconventional results: the spontaneous emergence of a recursive, infinite but learnable language. However, the learning algorithm used is very brittle, and it's difficult to extend the model to domains with more diverse semantics and a more heterogeneous syntax.

In contrast, Batali's model is very difficult to understand. It also uses a form of predicate logic to represent semantics, but it uses "exemplars" as the basic representation of the grammar, and "argument maps" to guide the combination of exemplars into meaningful sentences. The results show the emergence of a complex language, with properties similar to case marking and subordinate clause marking in natural languages. The emergent languages are essentially infinite but nevertheless learnable (from meaning–form pairs). The learning algorithm is successful and robust in this complex domain presumably because of the redundancy it allows.

3.4 Some features of explicit induction models

Several other explicit induction models have been studied. They have in common that no uniformity of languages are assumed. Typically, individuals in these models start with an empty grammar and empty lexicon, and gradually add new rules and lexical items based on the received sentences and observed patterns. Individuals are, however, equipped with an invention procedure, such that they can generate new sentences when required.

Further, in these models learning is typically from form–meaning pairs and a lexicon is built-up in parallel with the grammar. The recognition of phonemes and the pragmatics of dialogs are built-in as assumptions of the models.

The models are all implemented as computer programs. Typically, the models are rather concrete: they consist of a population of individuals, with procedures for production, invention, interpretation and induction, and a set of possible message to communicate. The languages studied in these models are still relatively simple, and exhibit just some basic word orders or morphological markers for the semantic roles of agents, patients and action. Empirical data from historical linguistics has so far played no role in these studies.

4 Discussion

I have reviewed some models of language acquisition and language change from two different traditions. The crucial question – which approach is best? – is still largely open to discussion. The following issues are important in comparing both approaches:

Learnability - Theoretical arguments. From the field of learnability theory it has sometimes been argued that grammar induction is impossible. In section 3.1 I have argued that this position is based on a misunderstanding of the negative learnability results. Learnability, however, is an important test for the validity of a grammar formalisms and induction algorithms. The challenge is to find a combination of a formalism that is as expressive as human languages are (i.e. mildly context-sensitive), and a learning algorithm that can induce it from the available primary linguistic data. In my view, parameter setting

models meet this challenge, but only by making unsatisfactory assumptions on the prior knowledge the algorithms start with. Explicit induction models, on the other hand, present considerable progress (i.e. most work with context-free grammars), but more work still needs to be done.

Learnability - Empirical arguments. From the field of psycholinguistics it has been argued that children have prior knowledge of syntactic constructions, because they choose, from apparently many logical possibilities that are consistent with the received evidence, the correct, seemingly arbitrary option. Grammar induction models, in this view, are – if not impossible – irrelevant, because children do not do induction. I believe that explicit induction algorithms have already shown that the logic of this argument is false. There is no need for assuming explicit prior knowledge, because the outcome of the interaction between learning biases and training data is subtle and often unexpected. Moreover, because languages are transmitted culturally from generation to generation, seeming arbitrary choices are likely to be the correct ones, because previous generations have used the same arbitrary learning algorithm to learn their language (Deacon, 1997; Kirby, 2000; Briscoe, 2002a; Zuidema, 2003).

Equivalence More subtly, it has been suggested that explicit induction models might in some sense be equivalent to parameter setting models. If the space of grammars that induction algorithms explore is finite, then that space could in principle be parametrized and hence described by a finite number of parameters. The induction algorithm can then be described, albeit possibly in a clumsy and complicated way, as a parameter setting procedure. If this is true – and it presumably is for the context-free grammar and finite-state machine inducers – the crucial issue is parsimony and clarity. Presumably, for some purposes the representation in terms of parameters is more useful, but for comparison with psycholinguistic, neurological and historical data the explicit grammar representation seems more appropriate. Further, the parameterized representation leads naturally to the uniformity assumptions, whereas the explicit grammar representation leads naturally to the view that grammars grows over time. Finally, stochastic grammar formalisms can not be parametrized in the concise way that parameter setting models usually assume. Worse, lexicalized, exemplar-based models can not be parametrized because there are infinitely many probability distributions that can be assigned to the string set (Bod, 1998).

In conclusion, the two approaches to modeling of language change are rooted in different theoretical positions on the nature of language and language acquisition. If one adopts the Principles and Parameters framework, the parameter change approach is the appropriate way to conceptualize language change. However, this approach requires more work to make explicit how each parameter is to be interpreted, which triggers for each parameter are available, how the child learns her lexicon and recognizes syntactic categories in the sentences it receives, how parameters depend on each other, etc. Moreover, it requires a satisfactory explanation for the evolution and development of the Universal Grammar in the child's brain. However, some Explicit Induction models might, even if one adopts this approach, still be useful as an equivalent representations that can be more easily compared to empirical data.

If one rejects the Uniformity Hypothesis and conceptualizes grammar acquisition as the gradual built-up of a grammar in the mind of the child, explicit induction models are the appropriate approach. Parameter change models are still useful as simple, but mathematically sophisticated models of how conventions spread in a population.

References

ANGLUIN, D. (1980). Inductive inference of formal languages from positive data. *Information and Control* **21**, 46–62.

BATALI, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe (2002b).

BERTOLO, S., ed. (2001). *Language Acquisition and Learnability*. Cambridge University Press.

BOD, R. (1998). *Beyond Grammar: An experience-based theory of language*. Stanford, CA: CSLI.

BRISCOE, T. (2002a). Grammatical acquisition and linguistic selection. In: Briscoe (2002b).

BRISCOE, T., ed. (2002b). *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

DEACON, T. (1997). *Symbolic species, the co-evolution of language and the human brain*. The Penguin Press.

GOLD, E. M. (1967). Language identification in the limit. *Information and Control (now Information and Computation)* **10**, 447–474.

HARE, M. & ELMAN, J. (1995). Learning and morphological change. *Cognition* **56**, 61–98.

HAUSER, M., CHOMSKY, N. & FITCH, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579.

HURFORD, J. R. (2002). Expression / induction models of language. In: Briscoe (2002b).

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

KIRBY, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford University Press.

KIRBY, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: *The Evolutionary Emergence of Language: Social function and the origins of linguistic form* (Knight, C., Hurford, J. & Studdert-Kennedy, M., eds.). Cambridge, UK: Cambridge University Press.

KIRBY, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe (2002b).

KIRBY, S. (2002b). Natural language from artificial life. *Artificial Life* **8**, 185–215.

KOMAROVA, N., NIYOGI, P. & NOWAK, M. (2001). The evolutionary dynamics of grammar acquisition. *J. Theor. Biology* **209**, 43–59.

LANGLEY, P. & STROMSTEN, S. (2000). Learning context-free grammars with a simplicity bias. In: *Proceedings of the Eleventh European Conference on Machine Learning*, pp. 220–228. Barcelona: Springer-Verlag.

MACWHINNEY, B., ed. (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.

NIYOGI, P. (1998). *The informational complexity of learning*. Boston, MA: Kluwer.

NIYOGI, P. & BERWICK, R. C. (1995). The logical problem of language change. Tech. rep., M.I.T.

NOWAK, M. A., KOMAROVA, N. & NIYOGI, P. (2001). Evolution of universal grammar. *Science* **291**, 114–118.

STEELS, L. (1999). The puzzle of language evolution. *Kognitionswissenschaft* **8**.

STOLCKE, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley.

WEXLER, K. & CULICOVER, P. (1980). *Formal principles of language acquisition*. Cambridge MA: MIT Press.

WOLFF, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication* **2**, 57–89.

YANG, C. D. (2000). Internal and external forces in language change. *Language Variation and Change* **12**, 231–250.

VAN ZAANEN, M. (2001). *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, School of Computing, University of Leeds.

ZUIDEMA, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.). Cambridge, MA: MIT Press. (forthcoming).

A Memory-less learner and batch learner

To estimate the probability that memory-less learning finds the correct grammar after a certain number (b) of sample sentences, we need to consider the inverse: the probability that the algorithm still has a wrong hypothesis after b sample sentence.

$$P(right\ grammar\ after\ b\ samples) = 1 - P(wrong\ grammar\ after\ b\ samples) \tag{1}$$

The probability that the learner still has the wrong hypothesis, depends on the probability that it initially chose the wrong hypothesis (simply $(N-1)/N$) times the probability that it remained for all b sentences at a wrong hypothesis. If it makes no essential difference which wrong grammar is the present hypothesis and how long it has held it as the hypothesis[5], the probability that the algorithm remain for b sentences at a wrong hypothesis is simply $P(remain)^b$. Hence,

$$P(wrong\ grammar\ after\ b\ samples) = \frac{(N-1)}{N}\left(P(remain)\right)^b \tag{2}$$

The probability to remain at a wrong grammar for each random input sentence is given by the probability that that input sentence happens to be consistent with the present (wrong) grammar, plus the probability that the algorithm jumps to another wrong grammar:

$$P(remain) = P(consistent) + P(another\ wrong\ grammar) \tag{3}$$

The probability that a sentence is consistent with a wrong grammar is simply the similarity parameter a in Nowak *et al.* (2001). The probability that the algorithm jumps to another wrong grammar is given by the probability that the input sentence is inconsistent $(1-a)$ times the fraction of other wrong grammars $((N-2)/N)$.

Putting all this together, the probability (q) that the memory-less learner has found the correct grammar after b input sentences is given by (Komarova *et al.*, 2001)[6]:

$$\begin{aligned} q_{memoryless} &= 1 - \frac{(N-1)}{N}\left(a + \frac{(N-2)(1-a)}{N-1}\right)^b \\ &= 1 - \frac{(N-1)}{N}\left(1 - \frac{(1-a)}{N-1}\right)^b \end{aligned} \tag{4}$$

The probability that the batch learner has found the correct grammar after b input sentences is found by Nowak *et al.* (2001) to be

$$q_{batch} = \frac{\left(1 - \left(1 - a^b\right)^N\right)}{(Na^b)} \tag{5}$$

[5]That is the case, for the memory-less learner, under the assumption of Nowak *et al.* (2001) that all grammars are equally similar to each other. In contrast, in a Principles & Parameters model, we can calculate the expected similarity based on estimates of how many parameters are revealed in a single sentence. Under the assumption that every sentence reveals m parameters, that all parameters are Boolean and that all parameters are revealed with equal probability: $a \approx \left(\frac{1}{2}\right)^m$. $a \approx \left(\frac{1}{2}\right)^m$. a is then an expected value rather than a constant, and equation (2) needs to be adapted. For simplicity, we will here follow the assumption of Nowak et al.

[6]Note that there is an error in this equation in Nowak *et al.* (2001) that is corrected in Komarova *et al.* (2001)

Phonemic Coding: Optimal Communication Under Noise?

Bart de Boer
Artificial Intelligence Lab
Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
bartb@arti.vub.ac.be

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute of Animal, Cell and Population Biology
University of Edinburgh
40, George Square
Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

http://arti.vub.ac.be/ ~ bartb
http://www.ling.ac.uk/ ~ jelle

Abstract

Human languages are universally phonemically coded, whereas many animal signal systems are not. A number of theories and models have been developed to explain this evolutionary transition, but some major problems remain. We present a simulation to investigate the hypothesis that phonemic coding is an side effect of optimizing signal systems for success in imitation. Crucially, signals in our model are trajectories in an (abstract) acoustic space. Hence, both holistic and phonemically coded signals have a temporal structure. Using both qualitative inspection of emerged systems of trajectories and a statistical analysis of a measure of phonemicity, we find that phonemically coded systems are indeed preferred. The model thus provides a new explanations for the evolutionary pathway to the emergence of phonemic coding.

1 Introduction

One of the universal properties of human language is the fact that it is phonemically coded. Linguistic utterances can be split into units that can be recombined into new linguistic utterances. For instance, the words "we", "me", "why" and "my" as pronounced in standard British English are built-up from the units "w", "m", "e" and "y", which can all be used in many different combinations.

There is some controversy about the exact level at which combination takes place. In the traditional view the atomic units are phonemes: minimal speech sounds that can make a distinction in meaning. An increasingly popular alternative view is that the atoms are syllables, or the possible onsets, codas and nuclei of syllables. Nevertheless, there is general agreement that in natural languages, atomic units are combined into larger wholes. For the purposes of this paper, we do not need to take

sides in the debate about the exact nature of the combinatorial elements of human language. Instead, we study signals that occur in an abstract acoustic space, and address the question of why and how phonemically coded sets of signals have emerged.

The combinatorial nature of human speech is in contrast with many animal calls and non-linguistic human utterances, which generally cannot be split into smaller units. The songs of some songbirds and whales, however, do seem to have combinatorial structure. The fact that in evolutionary unrelated lineages combinatorial systems have emerged indicates that such systems can be considered as evolutionary attractors. Recombination apparently has major evolutionary advantages. Two views on the advantages that recombination offers are:

1. It makes it possible to transmit an infinite number of messages over a noisy channel (the "noisy coding argument", an argument from information theory, e.g. Nowak & Krakauer 1999).

2. It makes it possible to create an infinitely extensible set of signals with a limited number of building blocks. Such productivity provides a solution for memory limitations, because signals can be encoded more efficiently, and for generalization, because new signals can be created by combining existing building blocks (the "productivity argument", a point often made in the generative syntax tradition, e.g. Jackendoff 2002);

These advantages are a good starting point for answering the questions of *why* combinatorial coding would emerge, and *how* initially holistic systems (which seem to be the default for smaller repertoires of calls) can change into phonemically coded systems. In this paper we will address both questions. In the following we will discuss some existing formal models of phonemic coding, discuss which open problems remain and then develop a model of our own that addresses some of these problems.

2 Previous work

2.1 Natural selection for combinatorial phonology

Several mathematical and computational models have shown that under noisy transmission, digital, combinatorial coding is more efficient than continuous coding. Nowak & Krakauer (1999) apply this insight in the context of the evolution of language, and derive an expression for the "fitness of a language". Imagine a population of individuals that all agree on which signals to use for which objects or events. The fitness of a language is now given by the expected success of a random individual to communicate about a random object or event with a random other individual. Nowak et al. show that when communication is noisy and when just a single sound is used for every meaning, the fitness is limited by an "error limit": only a limited number of sounds can be used — and thus a limited of meanings be expressed — because by using more sounds the successful recognition of the current signals would be impeded. Nowak et al. further show that in such noisy conditions, fitness is higher when (meaningless) sounds are combined into longer words. When the environment is combinatorial (i.e. objects and actions occur in many combinations) the fitness is highest when meaningful words are combined into longer sentences.

These results are essentially particular instantiations of Shannon's more general results on "noisy coding" (Shannon, 1948), as is explored in a later paper by the same group (Plotkin & Nowak, 2000). More interesting is the question how natural selection could favor a linguistic innovation in a population where that innovation is still very rare. Nowak & Krakauer (1999) do a game theoretic analysis of

"compositionality". They consider all mixed strategies where both holistic and compositional signals are used, and show that strategies that use more compositionality can invade strategies that use less. This means that the adaptive dynamics of languages under natural selection should lead to compositionality. For combinatorial phonology a similar analysis can be given.

Although this model is a useful formalization of the problem and gives some important insights, as an explanation for the evolution of phonemic coding and compositionality it is still insufficient. The main problem is that the model only considers the advantages of combinatorial strategies, and ignores two obvious disadvantages: (1) by having a "mixed strategy" individuals have essentially two languages in parallel, which one should expect to be costly because of memory and learning demands and additional confusion; (2) combinatorial signals that consist of two or more sounds take longer to utter and are thus more costly. A fairer comparison would be between holistic signals of a certain duration (where repetition of the same sound decreases the effect of noise) and combinatorial signals of the same duration (where the digital coding decreases the effect of noise). This is the approach we take in this paper.

2.2 Crystallization in the perception–imitation cycle

A completely different approach to phonemic coding is based on "categorical perception". Categorical perception (Harnad, 1987) is the phenomenon that categorization influences the perception of stimuli in such a way that differences between categories are perceived as larger and differences within categories as smaller that they really are (according to an "objective" similarity metric). For instance, infants already perceive phonemes as closer to the closest prototype phoneme from their native language than it is according to an "objective" (cross-linguistic) acoustical metric (Kuhl *et al.*, 1992). Hence, when presented vowels as stimuli ranging from /o/ to /a/ in fixed increments, British subjects will hear the first stimuli as o's or almost o's, and the last as a's or almost a's. Apparently, the frequency and position of acoustic stimuli gives rise to particular phoneme prototypes, and the prototypes in turn distort the perception.

Oudeyer (2002) studies a model that yields such a perceptual distortion. In this model, signals are modeled as points in an acoustic space, and are thus instantaneous. Oudeyer considers that signals survive from generation to generation because they are perceived and imitated. Oudeyer shows that categorical perception *shapes* a signal repertoire such that it conforms more and more to the prototype phonemes. Thus, emitted signals shape perception, and distorted perception shapes the repertoire of signals in the cycle from emission to perception to emission (the perception–imitation cycle; see also Westermann 2001 for a model of sensori-motor integration and its relevance for imitation and categorical perception). Oudeyer calls the collapse of signals in a small number of clusters "crystallization".

Oudeyer's model is fascinating, because it gives a completely non-adaptive mechanism for the emergence of phonemic coding. However, it is not clear how well it would work if signals have a time structure rather than being instantaneous[1]. Moreover, even if the mechanism works also in these conditions, it remains an important question whether phonemic coding increases the functionality of the language, and thus the fitness of the individual that uses it. If not, one would expect selection to work against it. In particular, in Oudeyer's model, where signals are instantaneous, a large repertoire of signals is collapsed into a small number of clusters. A functional pressure to maintain the number of distinct signals would thus have to either reverse the crystallization, or combine signals from different clusters. This aspect, which seems the core issue in understanding the origins of phonemic coding, is

[1] Oudeyer has also tested the model for sequences of sounds (Oudeyer, p.c.), but, as far as we know, not for continuous trajectories. It seems that in this version of the model the "combinatorial" aspects of phonemic coding is imposed and only the "categorical" (see section 2.3) aspect is emergent, such that our criticism still holds.

not modeled by Oudeyer. In our model, we ensure that the number of distinct signals remains at least at the same level; i.e. the functionality increases rather than decreases.

2.3 Aspects of phonemic coding

Other models of phonemic coding assume the sequencing of phonetic atoms into longer strings as given. They concentrate rather on the structure of the emerged systems (Lindblom *et al.*, 1984; de Boer, 2001; Redford *et al.*, 2001) or on how conventions on specific combinatorial signal systems can become established in a population through cultural transmission (Steels & Oudeyer, 2000). These models are interesting, and, importantly, bridge the gap with empirical evidence on how phonemic coding is implemented in the languages of the world.

It appears from this discussion that there are 4 related, but distinct aspects to phonemic coding:

1. Phonemically coded systems are *categorical*, in that they allow only a small number of basic sounds and not all feasible sounds in between;
2. they are also *superficially combinatorial*, in that all parts of each signal overlap with parts of other signals;
3. they are also *productively combinatorial*, in that the cognitive mechanism that produces and interprets signals uses the common parts of signals as building blocks that can be combined in all sorts of combinations;
4. the possible sets of categories and combinatorial rules show particular (cross-linguistic) constraints.

These aspects form a hierarchy, where the aspects further down the list imply the aspect above it. Oudeyer (2002) shows a non-adaptive mechanism that can yield aspect 1 (and gives a starting point for 4), but does not explain how the other aspects come about and how the functionality of the signal system is preserved. Nowak & Krakauer (1999) show how natural selection could favor 2, but ignore the temporal aspects of holistic signals. Zuidema & Hogeweg (2000) and Zuidema (2003) can be viewed as assuming aspects 1 and 2, and addressing the emergence of aspect 3 under natural selection and cultural evolution respectively (but the models are not discussed in these terms). Lindblom *et al.* (1984); de Boer (2001); Redford *et al.* (2001); Steels & Oudeyer (2000) all address aspect 4.

The question thus remains open as to under what circumstances a system of holistically coded signals with finite duration would change into a phonemically coded system of signals. In the paper we study a single mechanism that can yield aspects 1 and 2.

3 The model

In our model, we do not assume combinatorial structure, but rather study the gradual emergence of phonemic coding from initially holistic signals. We do take into account the temporal structure of both holistic and phonemically coded signals. We view signals as continuous movements ("gestures", "trajectories") through an abstract acoustic space. We assume that signals can be confused, and that the probability of confusion is higher if signals are more similar, i.e. closer to each other in the acoustic space according to some distance metric. We further assume that a functional pressure on distinctiveness maximizes the distance between trajectories.

3.1 Representing trajectories

The model is based on part-wise linear trajectories in a bounded 2-D continuous space (of size 15.0 × 15.0 in all simulations reported here). Trajectories are sequences of points with fixed length (here: 20). Each point has a fixed distance of 1.0 to the immediately preceding and following points in the sequence. The following and preceding points to a point can lay anywhere on a circle of radius one with that point at the center. Trajectories always stay within the bounds of the defined acoustic space.

Signals in the real world are continuous trajectories, but in the model we need to discretize the trajectories. However, to ensure that we do not impose the phonemic structure we are interested in, we discretize at a much finer scale than the phonemic patterns that will emerge. Hence, the points on a trajectory are not meant to model atomic units in a complex utterance.

3.2 Measuring distances

The distance between two trajectories t and r is defined as the sum of the distances between all corresponding points in the best possible alignment of the two trajectories. In finding the best possible alignment, one point from t can be mapped on several neighboring points in r and vice versa. In this way trajectories that resemble each other in shape, but that do not align perfectly still are considered close. This models the way humans perceive signals. The distances are calculated using "dynamic time warping", an efficient method that before the advent of statistical models, has been used with reasonable success in computer speech recognition (e.g. Sakoe & Chiba, 1978).

3.3 Maximizing the total mutual distance

In the first set-up of the model, we consider an idealized single repertoire of trajectories that, in a sense, repel each other. That is, the total distance between trajectories is optimized using a simple hill-climbing algorithm. The model goes through a large number of iterations. At every iteration, the sum of all mutual distances is calculated. Then a random change is applied to a random trajectory t, and the total distance is measured again. If this second measurement is larger than the first, the change is kept. If not, the change is reverted.

Random changes always respect the constraints on well-formed trajectories. Hence, a random point, t_x, is moved to a new random position (from a Gaussian distribution around the old position, provided it falls within the boundaries of the acoustic space). The two points on both sides of the moved point, t_{x+1} and t_{x-1}, are moved closer or further away such that the distance to t_x is again 1. The direction from t_x to t_{x+1} or t_{x-1} remains the same, unless the point would cross the boundary of the space, in which case it is rotated to the closest point within the boundary at distance 1 from t_x. The same procedure is applied recursively to the neighbors of t_{x+1} and t_{x-1} until the ends of the trajectory are reached.

In the second set-up of the model, we investigate what kind of repertoires of trajectories emerge in a *population* of agents that try to imitate each other in noisy conditions. The model is very similar, but now each agent in the population has its own repertoire, and it tries to optimize its own success in imitating and being imitated by other agents of the population.

This version of the model is like the imitation games of de Boer (2000). These only modeled holistic signals (vowels) and did not investigate phonemic coding. The game implemented here is a slight simplification of the original imitation game. First, all agents in the population are initialized with a random set of a fixed number of trajectories. Then for each game, a speaker is randomly selected from the population. This speaker selects a trajectory, and makes a random modification to it. Then it plays a number of imitation games (50 in all simulations reported here) with all other agents in

the population. In these games, the *initiator* utters the modified trajectory with additional noise. The *imitator* finds the closest trajectory in its repertoire and utters it with noise. Games are successful if the imitator's signals is closest to the modified trajectory in the initiator's repertoire. If it turns out that the modified trajectory has better imitation success than the original trajectory, the modified trajectory is kept, otherwise the original one is restored.

4 Results

4.1 Optimizing a single repertoire

We ran the model under the single repertoire condition with a number of different parameters. In all simulations the initial trajectories are random sequences of positions, where the only constraints are that neighboring points are at distance 1 from each other and that all points are within the permitted space.

In simulations with few trajectories (up to 4), we find that the trajectories "bunch-up" and remain within a very small area at maximum distance from the areas used by the other trajectories. Each of these signals is thus a holistic signal, but the signals are "categorical".

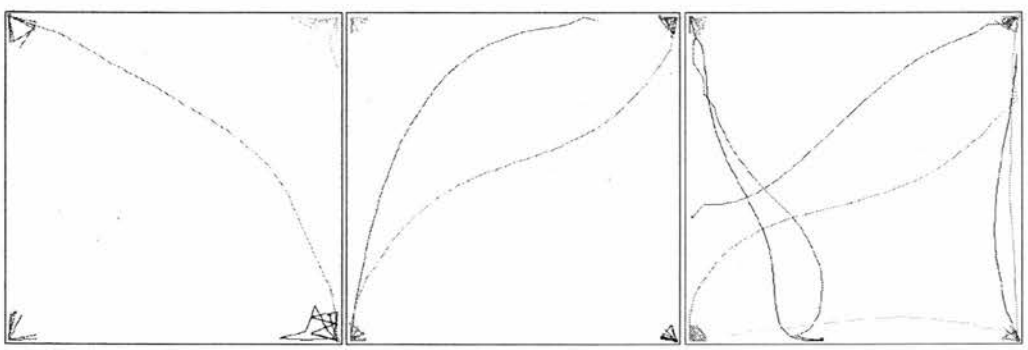


Figure 1: Comparison of optimized systems of 5, 6 and 10 trajectories. Note the reuse of start- and endpoints (squares indicate start points).

In simulations with 5 trajectories, 4 occupy the corners areas of the acoustic space in the same way as in simulations with 4 trajectories. However, the fifth trajectory stretches from one corner to another, and thus shares the areas for its begin and end points with two different other trajectories (see fig. 1, leftmost panel). This can be interpreted as a rudimentary phonemic code.

With more trajectories, the reuse of beginning and end points becomes more pronounced. In the simulation with 6 trajectories, the first 5 are similarly organized, but the sixth is essentially the inverse of the fifth. In the simulation with 10 trajectories, 3 trajectories are still bunched up in a small area of the acoustic space, but the other 7 are stretched out, sharing begin and end points with one another. Frequently one can find trajectories that are more or less the inverse of another trajectory.

In order to perform a statistical analysis, a numerical measure of the extent to which emerged systems were phonemically coded had to be defined. This measure, called the phonemicity $\mathcal{P}$, is defined as the ratio between the average distance between the start and end points of all trajectories and the average distance between all other corresponding points of all trajectories. Corresponding points are defined as points that are an equal number of steps away from either a start or an end point

(i.e., two points that are at position 3 are corresponding points, but so are a point at position 3 and position L-2). The details for this measure are in the appendix.

In a phonemically coded system of trajectories, start and end points are expected to be closer together than the other points on a trajectory, while in a holistically coded system of trajectories, the average distance is expected to be approximately equal. Therefore, the measure should give lower values for phonemically coded systems. It is quite likely that better measures of phonemicity can be defined, but this measure does make a distinction between holistically and phonemically coded systems, and was therefore adopted for the analysis.

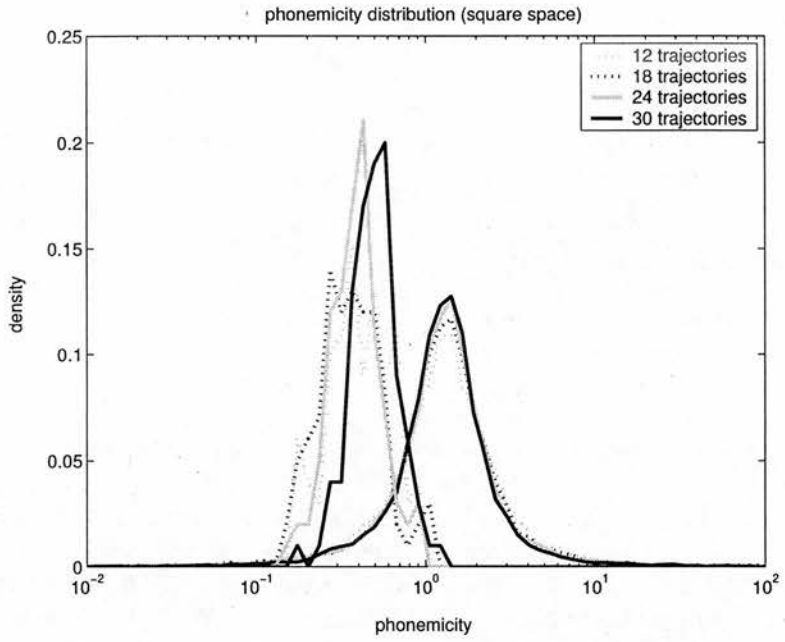


Figure 2: Distributions of phonemicity of random systems (right peak) and optimized systems (left peak). Note that the phonemicity measure for optimized systems is lower, indicating that the optimized systems are more phonemically coded than random systems.

Two conditions were compared. In both conditions the systems of trajectories were initialized randomly; only in the second condition were systems of trajectories optimized for distance first using 30,000 optimization steps. The results were measured for systems of many different sizes, but are presented for systems of 12, 18, 24 and 30 trajectories in figure 2. 10,000 random systems were evaluated, but for computational reasons only 100 optimized systems, as the amount of computation needed for optimization precluded larger numbers of systems to be evaluated. Note that the horizontal axis (showing the phonemicity) is logarithmic. This has the advantage of both making the peaks more distinct and making the distributions more similar to the normal distribution. When using the t-test, on both the phonemicity and its logarithm, it turns out the difference between the distribution of the random systems and the optimized systems is significant with $p < 0.05$ (the t-test is less appropriate for the non-log measure, because of the highly skewed distribution).

This result indicates that optimization for acoustic time-warped distance between trajectories results in more phonemically coded systems.

4.2 Optimizing repertoires in a population

For vowel systems, it has been shown that optimizing a single repertoire leads to similar systems as a population-optimization system (compare de Boer, 2000; Liljencrants & Lindblom, 1972). It can be shown that for trajectories the same is true, under the condition that noisy distortions of trajectories do not distort the shape of these trajectories too much. This is illustrated in figure 3.

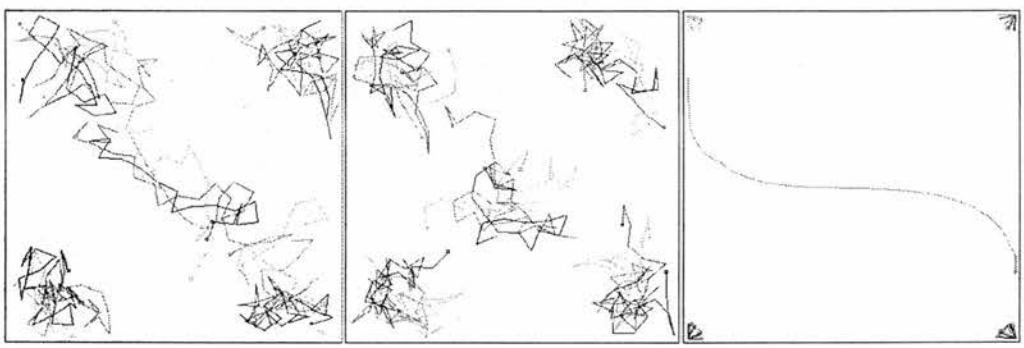


Figure 3: From left to right: emerged system with five trajectories in a population of ten agents (four agents shown), emerged system with five trajectories and uncorrelated noise, and optimized system of five trajectories. Small squares indicate the starting point of trajectories.

In this figure the left frame shows the system of five trajectories that resulted from playing imitation games in a population, using form-preserving noise. The right frame shows a system of five trajectories that resulted from optimizing distance. It can be observed that in both cases, the corners are populated by four trajectories, which are bunched up. The fifth trajectory, in contrast, follows the diagonal. As before, an analysis in terms of phonemes suggests itself: the four corners are basic phonemes, while the fifth trajectory uses one as the corners as a starting phoneme and the opposite corner as the ending phoneme. Both models result in similar systems of trajectories.

The middle frame, on the other hand, shows that when noise does not preserve shape of trajectories, a system results in which all trajectories are bunched up and an analysis in terms of phonemes is therefore not possible. As noise in real signals is band limited, it follows that shape will always be preserved to some extent. Therefore the shape-preserving model is indeed the correct model. Instead of investigating computationally extremely costly population models, it is therefore possible to investigate emergence of phonemic coding using the optimization model. For computational reasons, we have not performed simulations in the population condition with more than 5 trajectories.

5 Conclusion

We have investigated whether systems of trajectories that are used for imitation in a population would tend toward phonemic coding when agents tried to maximize their imitation success. It was found that running simulations of populations directly was too time-consuming. However, it was also found that direct optimization of time-warped distance between trajectories resulted in systems of trajectories that were similar to those found in preliminary experiments with imitation in a population. For this to be true, it was necessary to assume that in the population case, shape of trajectories was preserved under noise. This is a realistic assumption, as it turns out to be true for all noise that is band-limited, i.e. for which the energy of higher frequencies tends to zero. This is the case for all real-world noise.

When systems of trajectories were optimized for time-warped distance, it turned out that start- and endpoints were reused and that there were no trajectories (at least for limited numbers of trajectories) that had the same start- and endpoint and that only differed in the shape of the trajectory in between. This is indicative of phonemic coding. A measure of phonemicity was defined and it was found that optimized systems had significant lower values for this measure than random systems, indicating that they were more phonemically coded.

The conclusion to be drawn from this is that systems of complex articulations (trajectories) that have maximum distance to each other tend to show aspects of phonemic coding. Systems that have trajectories that are maximally distant from each other are most robust to noise. This means that optimizing systems of large numbers of complex articulations for robustness to noise, which is likely to happen when they are used for communication in a population, would result in systems of trajectories that can be analyzed in terms of phonemes.

The relevance for the evolution of speech is clear. When populations of agents start to communicate using small numbers of signals, it is unlikely that they would use phonemic coding, or be able to use it if it occurred. However, when extending the number of signals, the most robust systems would be the ones that can be analyzed as phonemically coded. Agents that have adaptations to detect and use this property would have an evolutionary advantage, as they would be able to learn faster, and probably to communicate more accurately as well. This provides a cultural beginning of a possible biological adaptation for using phonemically coded signals. This adaptation in the area of speech could later be exapted for use in combining words, in other words, for syntax.

6 Future work

The results described in this paper are preliminary, and need to be extended in several ways. Firstly, the model, especially in the population condition, should be tested with larger number of trajectories, and with trajectories of longer length. Presumably, the "phonemic coding" would then not just apply to the start and end points of the trajectories. Consequently, another measure of phonemicity needs to be defined.

Further, the model can be altered such that it allows trajectories of varying length in a single repertoire, and perhaps varying distances between the points of a trajectory.

Finally, and most ambitiously, the model should be extended to incorporate the aspects of phonemic coding that are currently not addressed: productive combinatorics and realistic constraints on the categories and rules of combination.

References

DE BOER, B. (2000). Self organization in vowel systems. *Journal of Phonetics* **28**, 441–465.

DE BOER, B. (2001). *The origins of vowel systems*. Oxford, UK: Oxford University Press.

HARNAD, S. (1987). *Categorical Perception*. Cambridge, UK: Cambridge University Press.

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

KUHL, P., WILLIAMS, K., LACERDA, F., STEVENS, K. & LINDBLOM, B. (1992). Linguistic experience alters phonetic perception in infants by 6 month of age. *Science* **255**, 606–608.

LILJENCRANTS, J. & LINDBLOM, B. (1972). Numerical simulations of vowel quality systems: the role of perceptual contrast. *Language* **48**, 839–862.

LINDBLOM, B., MACNEILAGE, P. & STUDDERT-KENNEDY, M. (1984). Self-organizing processes and the explanation of language universals. In: *Explanations for language universals* (Butterworth, M., Comrie, B. & Dahl, ., eds.), pp. 181–203. Berlin: Walter de Gruyter & Co.

NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.

OUDEYER, P.-Y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In: *Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior* (Hallam, B., Floreano, D., Hallam, J., Hayes, G. & Meyer, J.-A., eds.), pp. 406–416. Cambridge, MA: MIT Press.

PLOTKIN, J. B. & NOWAK, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology* pp. 147–159.

REDFORD, M. A., CHEN, C. C. & MIIKKULAINEN, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech* **44**, 27–56.

SAKOE, H. & CHIBA, S. (1978). Dynamic programming optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**, 43–49.

SHANNON, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal* **27**, 379–423 and 623–656.

STEELS, L. & OUDEYER, P.-Y. (2000). The cultural evolution of syntactic constraints in phonology. In: *Proceedings of the VIIth Artificial life conference (Alife 7)* (Bedau, M. A., McCaskill, J. S., Packard, N. H. & Rasmussen, S., eds.). Cambridge (MA): MIT Press.

WESTERMANN, G. (2001). A model of perceptual change by domain integration. In: *Proceedings of the 23d Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

ZUIDEMA, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.). Cambridge, MA: MIT Press. (forthcoming).

ZUIDEMA, W. & HOGEWEG, P. (2000). Selective advantages of syntactic language: a model study. In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (Gleitman & Joshi, eds.), pp. 577–582. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix: Measuring phonemicity

The average distance $\mathcal{E}$ between the extreme points (start and end points) is given by:

$$\mathcal{E} = \frac{1}{2N(N-1)} \sum_{i=0}^{N} \sum_{j=i+1}^{N} \left(D\left(i_1, j_1\right) + D\left(i_1, j_L\right) + D\left(i_L, j_1\right) + D\left(i_L, j_L\right) \right) \tag{1}$$

where N is the number of trajectories and L is the length of each trajectory. The function D is a measure of distance between points and will be explained below. The average distance between all other corresponding points is given by:

$$\mathcal{C} = \frac{1}{N(N-1)(L-2)} \sum_{i=0}^{N} \sum_{j=i+1}^{N} \sum_{k=2}^{L-1} \left(D\left(i_k, j_k\right) + D\left(i_{L-k+1}, j_k\right) \right) \tag{2}$$

The phonemicity $\mathcal{P}$ is then simply:

$$\mathcal{P} = \frac{\mathcal{E}}{\mathcal{C}} \tag{3}$$

The distance function $D(i_a, j_b)$ is the inverse, squared Euclidean distance between point a of trajectory i and point b of trajectory j:

$$D\left(i_a, j_b\right) = \frac{1}{\epsilon + |p_a(i) - p_b(j)|^2} \tag{4}$$

where $p_a(i)$ is the position of point a of trajectory i. The term ϵ ($\epsilon = 0.01$ throughout this paper) is added to avoid division by zero. Note that this is a different distance function than was used in the optimization of the distances between trajectories.

Mathematical Linguistics and Language Evolution

Timothy J. O'Donnell (timo@inf.ed.ac.uk) and Willem Zuidema
University of Edinburgh, Language Evolution and Computation Research Unit

Human languages exhibit a combination of computational features that make them unique systems of communication in nature: large and learned lexicons, combinatorial phonology, compositional semantics, and hierarchical phrase structure. In the field of evolution of language controversies have often focused on the complexity of these computational mechanisms. These controversies include debates about innateness, whether or not language was exapted, if it is the result of a few or many mutations and if it increased in complexity over evolutionary time (see e.g. Pinker & Bloom, 1990, and the many peer commentaries and the authors' response in the same issue). We analyze these debates and find that at their core they rely in varying degrees on two implicit assumptions: (i) that complexity in the computational machinery for processing language is difficult for evolution to achieve and/or that (ii) that complexity is itself a trait which can be selected for or against.

Out of the many possible ways of studying computational complexity, formal linguistics has primarily been concerned with situating natural language processes and formalisms on various computational hierarchies. By far the most studied of these is the (extended) Chomsky Hierarchy. We ask the questions: how do the two assumptions outlined above fare when analyzed under this notion of complexity, and how does this apply to the debates in the field? Such a formal definition would potentially resolve conflicting intuitions about complexity (exemplified e.g. in Lewontin's and Piatelli-Palmarini's commentaries on Pinker & Bloom, 1990).

We argue that complexity in the automata theoretic sense is in fact very common in natural systems. We find it plausible that genes can code for systems with small numbers of elements interacting with simple rules. There is increasing evidence that these sorts of systems are in fact often computationally universal (e.g. Wolfram, 2003). Furthermore, certain classes of neural network models have been shown to be Turing equivalent (Siegelmann & Sontag, 1991), and capable of efficiently encoding phenomena such as hierarchical phrase structure (Pollack 1988). We suspect that the reality is that brains in many kinds of animals are already implementing algorithms and computations which are sufficiently complex to represent and process language in the strict automata theoretic sense.

Furthermore, we go on to argue that these grammars and automata are not well suited to be used as phenotypes in biological models. They do, of course, expose interesting differences in grammatical classes on the hierarchy. For instance, the word recognition, or parsing problem increases in time complexity as one makes certain moves up the hierarchy. Likewise, differences in the hierarchy can be understood in terms of increasing relaxation of memory limitations, e.g. finite to stack based to stack based with less restrictive push procedures, etc. But it is difficult to see how these differences satisfy various evolutionary constraints or can affect fitness. We argue that instead of looking at these formalisms in terms of their place on the hierarchy we must look deeper at the properties of language that they

are meant to abstract.

We summarise that it is not the physical constraints of the general neural architecture that restrict natural language to a specific complexity class. Rather, the requirements of learnability and population coherence as well as the interface conditions of interpretability and producability under realistic time and noise constraints choose specific classes of computational mechanisms. These mechanisms restrict any language that is to survive either cultural or genetic transmission. We discuss the implications of this for the debates outlined in the introduction and reach some general conclusions. For instance, is it theoretically useful to describe the evolution of language as climbing the Chomsky hierarchy? (as do, e.g. Hashimoto & Ikegami, 1996). Finally, we conclude that while the Chomsky hierarchy is a bad model of phenotypic complexity, it is a very good model of language. This suggests a way of rescuing it as a tool for evolutionary theory.

Hashimoto, T. & Ikegami, T. (1996). The emergence of a net-grammar in communicating agents. BioSystems 38, 1-14.

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. Behavioral and brain sciences 13, 707-784.

Pollack, J. B. (1988). Recursive auto-associative memory: Decising compositional distributed representations. In: Proc. of the Tenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum.

Siegelmann, H. and Sontag, E. (1991). Neural networks are universal computing devices. Technical Report SYCON--91--08, Rutgers Center for Systems and Control.

Wolfram, S. (2002). A New Kind of Science. Champaign, IL, U.S.A.: Wolfram Media.